New York State Mortgage

I. Introduction

Mortgage loans play a significant role in shaping housing affordability, and we will explore their crucial influence in one of the largest and most economically diverse cities in the U.S. This project aims to explain the dynamics of mortgage loan amount to borrowers in the counties of New York state, U.S., that offered by the State of New York Mortgage Agency (SONYMA) with regards to their backgrounds such as household size, the type of property they desire and their financial background. Doing so, we will be able to provide someone who is looking for a mortgage loan in New York with SONYMA data-driven evidence on how their geolocation and personal background and their specific situations affect the amount of loan they may receive.

II. Data Analysis

IIa. Data Information

The data we are using is published by the State of New York under Open Data NY. It can be found on [https://data.ny.gov/Economic-Development/State-of-New-York-Mortgage-Agency-SONYMA-Loans-Pur/22ew-dxez/about_data]

The specific information about the variables being explored are as follows:

- **Original Loan Amount (Numeric)**: The response variable which is the original mortgage amount borrowed encoded in US dollars.
- **SONYMA DPAL/CCAL Amount (Numeric)**: the amount of DPAL/CCAL support offered by SONYMA encoded in US dollars.
- **Bond Series (Text)**: The bond series from which used by SONYMA to purchase the mortgage encoded as a categorical variable with 120 categories.
- **Loan Purchase Date (Text)**: The date the loan is purchased by SONYMA from the original mortgage lender.
- **Purchase Year (Numeric)**: The year the loan is purchased by SONYMA from the original mortgage lender.
- **Original Loan To Value (Numeric)**: The ratio of the original mortgage loan and the value of the house encoded as 0-100 ratio.
- **Loan Type (Text)**: The type of mortgage loan provided to the homebuyer encoded as a categorical variable with 2 categories.
- **Original Term (Numeric)**: The term length of the mortgage provided to the homebuyer in days.
- **County (Text)**: The county where the house being mortgaged is located.
- **FIPS Code (Numeric)**: The numeric equivalent to County.
- **Number of Units (Text)**: The number of units being funded by the mortgage.
- **Property Type (Text)**: The type of property being funded by the mortgage encoded as a categorical variable with 7 categories.
- **Housing Type (Text)**: The construction status of the house(s) being funded by the mortgage encoded as a categorical variable with 3 categories.
- **Household Size (Numeric)**: The number of person(s) living in the house funded by the mortgage.

Data Cleaning

The original dataset contains Nulls in the SONYMA amount with around 26.3% (7516/28528). Choosing the loan amount as the response variable, it significantly depends on the SONYMA amount because as stated in the overview of the dataset, the SONYMA amount column implicitly represent the borrower financial background: credits report, income information, bank statements, property appraisals, etc. The correlation

heat map below will show more detailedly other numeric explanatory will affect the response variable loan amount.

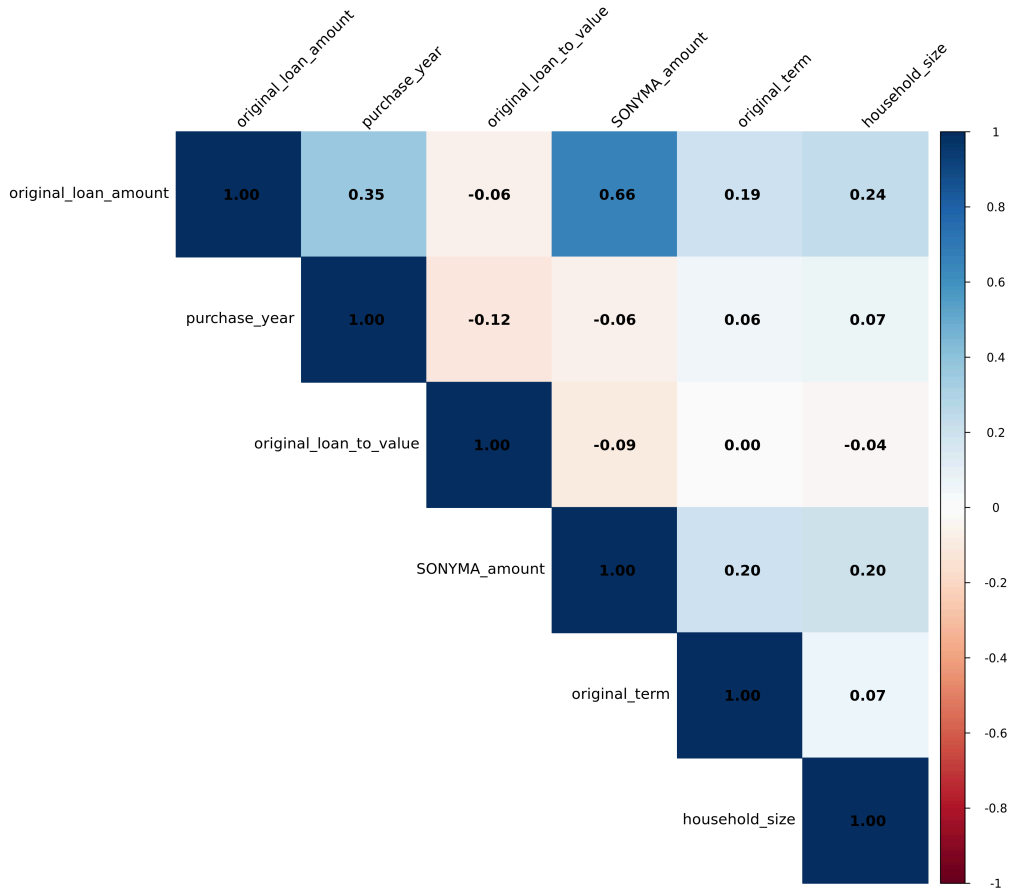
Exaplanatory Data Analysis

Data Selection

As examining the loan amount that a borrower can get based on their application and SONYMA additional form assessment. We believe that the "Bond Series" and "Number of Units" columns are extraneous because it relates to how SONYMA operates their organization (as the special purpose enity of the government to manage the mortgage responsibilities according to the overview file of the data can be found in the link above). Moreover, the date that the loan purchased by SONYMA is discarded because it is considered to be irrelavent since we will keep the year column which is more meaningful for our research. The Loan type also seems to be inessential because 98% of 28528 the transactions have the same type. We also dropping the FIPS Code column because it is the numeric representation of the County that is more rational to use as Text type.

Data Analysis

Collinearity Heatmap of All Variables (Figure 1)



The heat map shows the correlation of each numeric covariates against each other. The SONYMA amount has the highest correlation of 0.66 with the original loan amount as expected because it implicitly represents the financial background of the borrowers which is the most important criterion to evaluate and determine the amount of loan a borrower can access. Moreover, the year when the borrowers asking for the loan, their family size and original term seems also to affect the amount of loan they can get.

original_loan_amount and purchase_year: Moderate positive correlation of 0.35 – suggests a slight trend of higher loans being issued in later years.

Weak or No Relationships: Most of the other variable pairs have weak (close to 0) or no significant correlation:

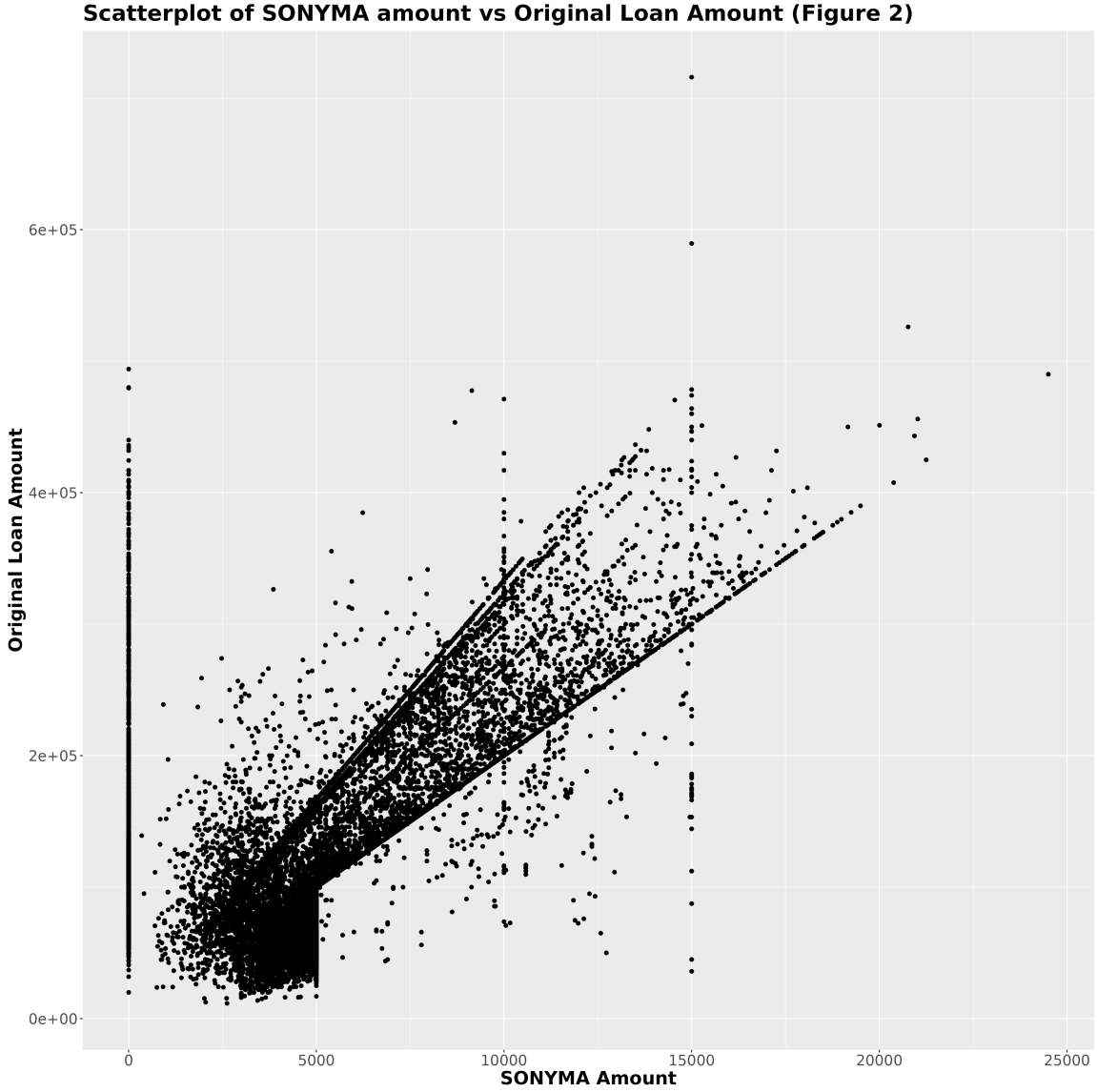
original_loan_to_value has very low correlation with most other variables.

purchase_year has minimal correlation with SONYMA_amount, original_term, and household_size.

household_size shows low correlations across the board (all < 0.25).

No Strong Multicollinearity Concerns:

The covariates are not highly correlated to each other so we can include all of them.



From the plot, we can see that there is a strong correlation between SONYMA amount and the original loan amount:

- SONYMA amount increase with the original loan amount as larger loan typically receives more SONYMA amount

Cluster near origin:

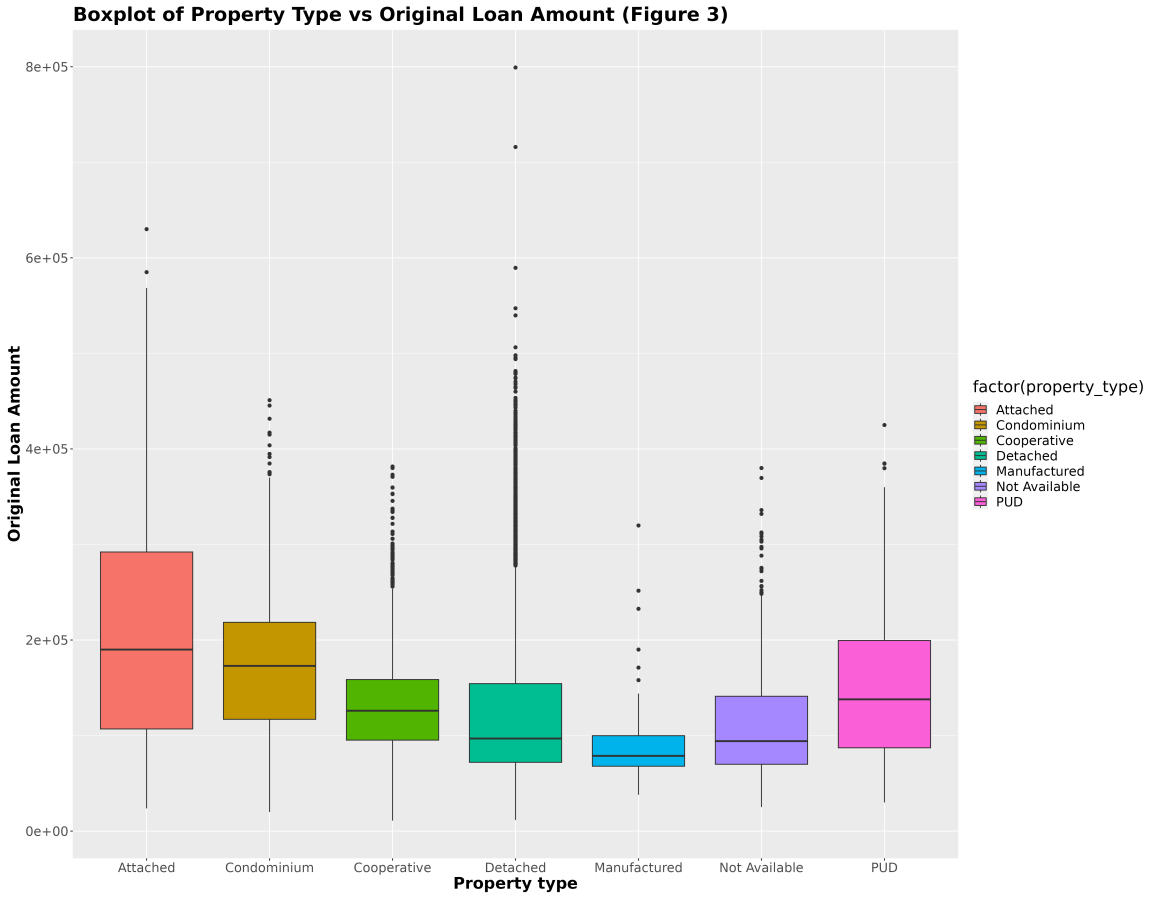
- There is a cluster of data points in the lower left corner (small SONYMA amount and small original loan amount), this suggest that many loans are received with little to no SONYMA amount and a high number of people have low mortgage amount

Vertical lines at SONYMA amount = 0:

- There's a thick vertical line at SONYMA amount = 0, meaning many borrowers received no SONYMA support regardless of their loan amount.
- Those points spread widely in the y-direction (loan amount), implying SONYMA doesn't always contribute to all loans.

The plot shows banding lines radiating out, which could indicate:

- Discrete levels of SONYMA support (e.g., fixed grant brackets).
- Rounding or policy-driven thresholds in support allocation.



Variation by Property Type:

The median loan amount differs across property types, suggesting property type plays a role in how much people borrow.

- Attached and PUD (Planned Unit Development) properties have some of the highest medians.
- Manufactured and Cooperative homes have the lowest medians, indicating smaller loan amounts.

Spread of Loan Amounts:

- Attached and Detached properties show a large spread, with long whiskers and many outliers—indicating more variability.
- Manufactured homes show much tighter boxes and smaller spreads, suggesting more consistent (and generally smaller) loan sizes since these homes are assembled on the factory off site then transported to the location, these homes typically are on the lower side of cost,.

Outliers:

- All property types include high-end outliers (especially Attached and Detached), some nearing $800,000.
- Cooperative and Manufactured types have fewer extreme outliers.

"Not Available" Group:

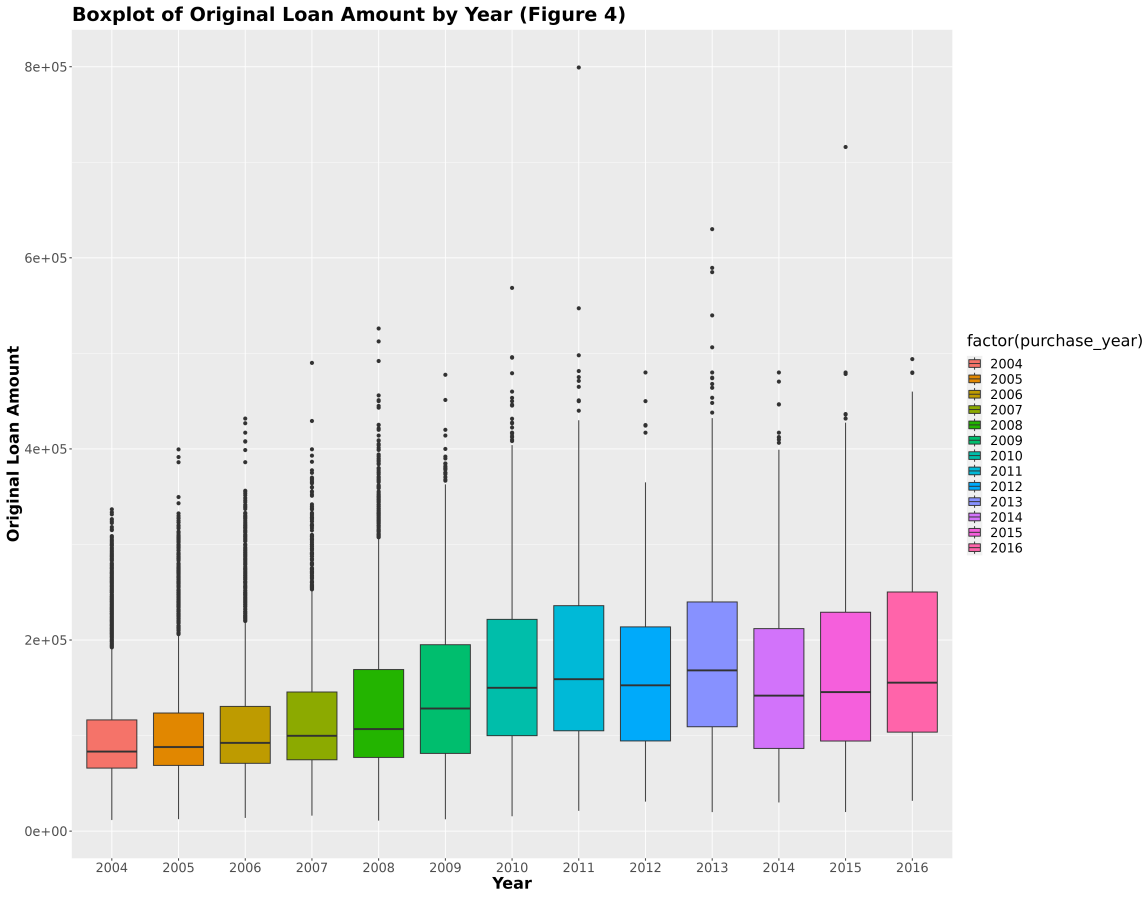
- This category has a wide spread and many outliers, which may indicate data quality issues or heterogeneous types lumped together.

PUD and Attached types not only have higher medians but also a larger upper quartile range—these may be associated with newer or higher-value developments.

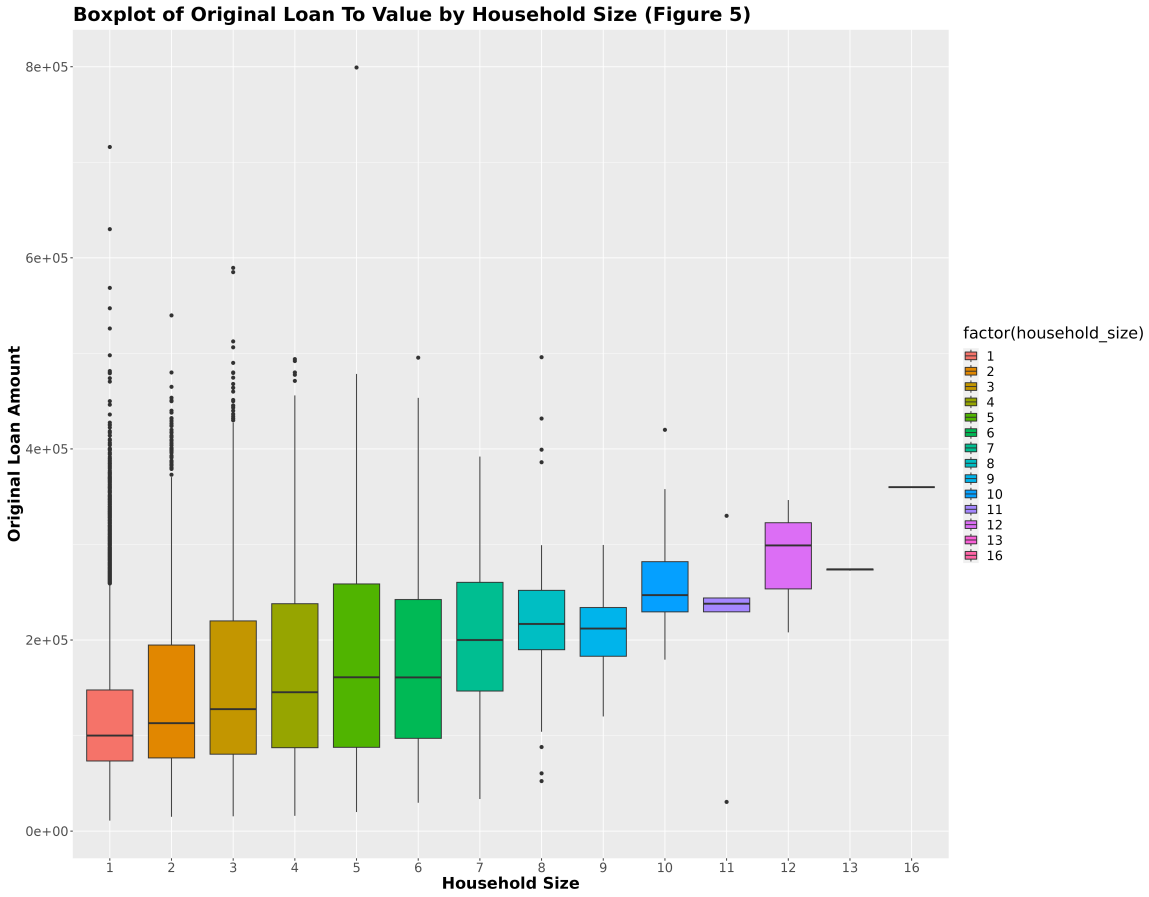
Possible Implications: Property Type as a Predictor: This categorical variable likely has predictive power

Segmentation: Borrower segments based on property type may behave differently in terms of loan size and affordability.

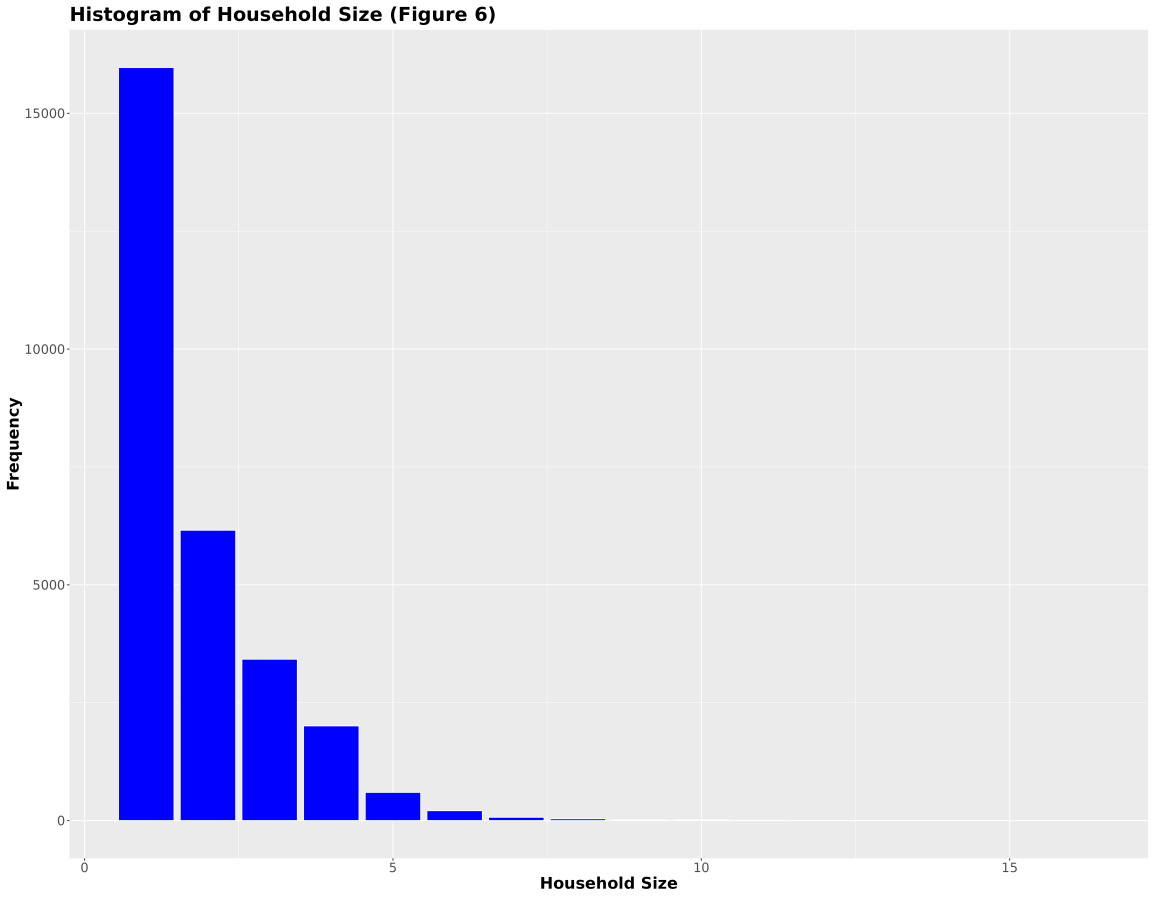
Policy Insight: Manufactured and cooperative home buyers may represent lower-income groups or more affordable housing options, potentially influencing funding strategies (e.g., SONYMA support).



The year that SONYMA purchased the loan to redistribute for the borrowers seems not directly affect the amount of money the borrower can negotiate but we believe that it implicitly does. The amount that SONYMA purchase in a year might implicitly reflect the economical situation of that year that will affect the number of borrowers as well as the amount of money and the inclination to start a mortgage loan. The plots show that the amount of purchasing steadily increase from 2004-2011 and fluctuate after that. We believe that overall, the tendency of buying new houses and requiring loan mortgage is increasing overtime and just fluctuate due to the financial crisis in the end of 2000s and in the beginning of the 2010s. However, in the long term, the amount of money people get as loan mortgage will increase.



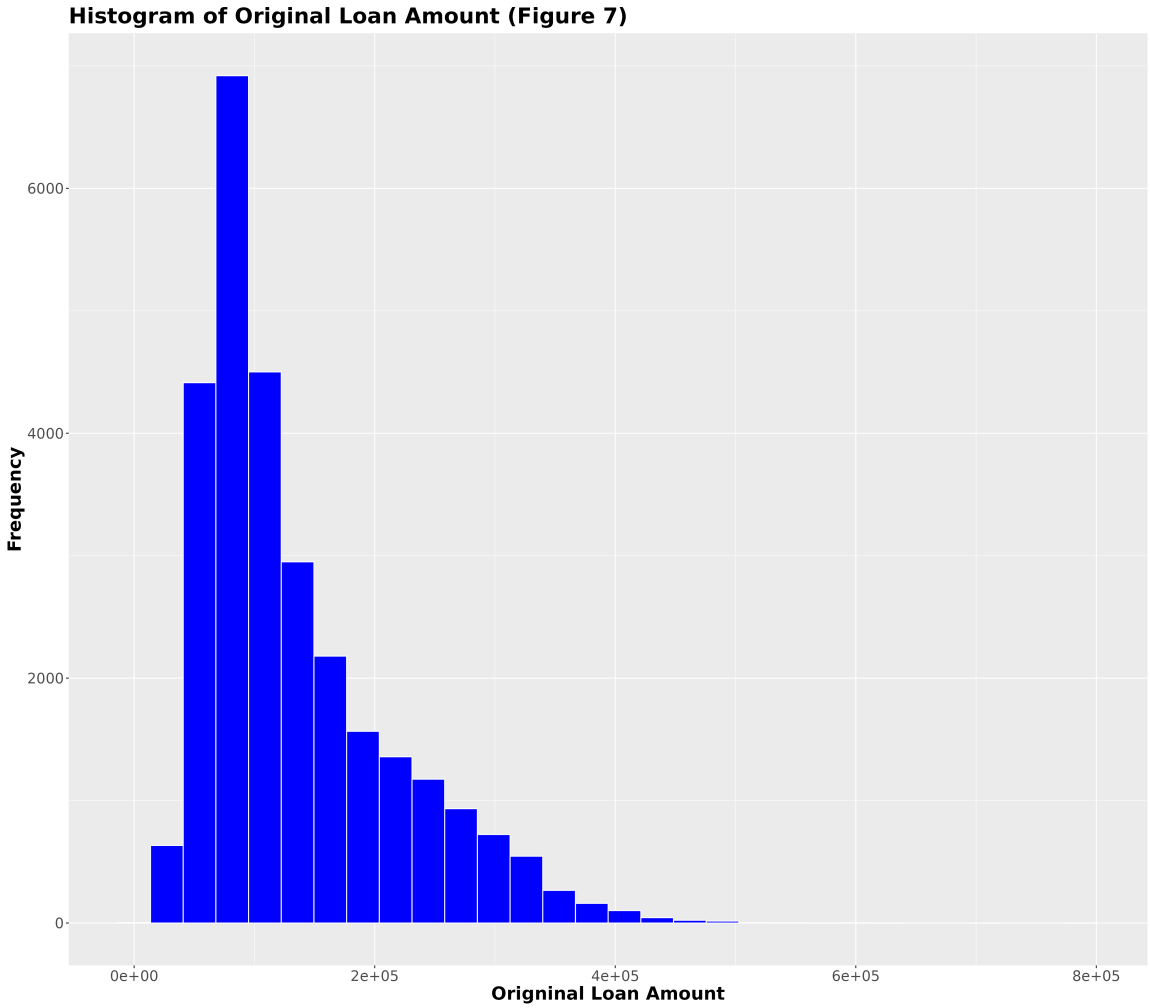
The plot suggests that households with larger size tend to have a greater loan amount. It can be explained that the larger the household size the larger space of the home they need and the more money in total they can make within a household. As a result, the loan amount they can get is larger to fulfill their demands.



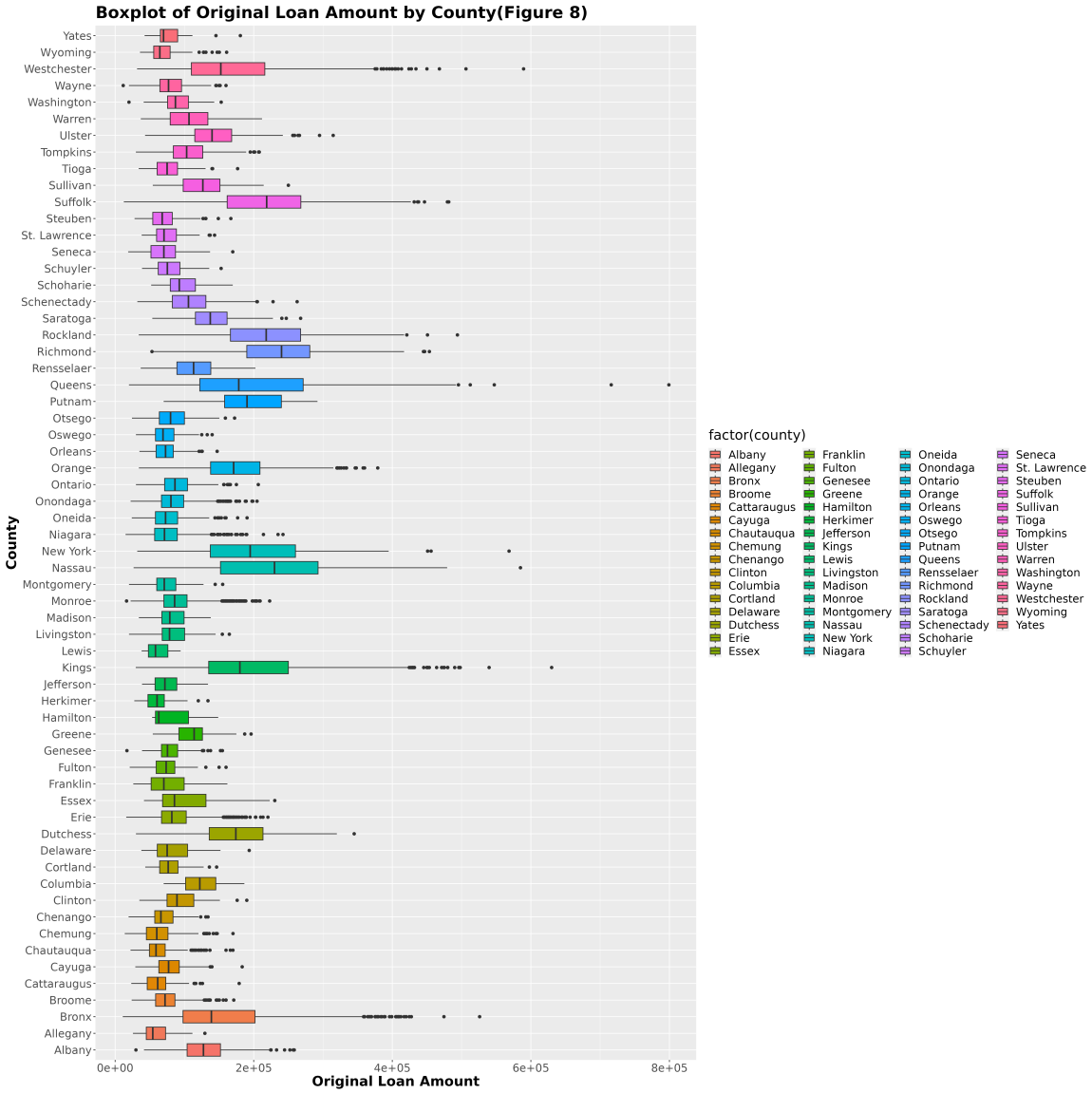
The most common household size of the borrower is 1 and 2. It is reasonable because most of them have low to moderate income and are first-time home buyer. They probably are the young people and young couples who have not formed a family yet and likely just are at the start of their careers and adulthoods.

Moreover, from this boxplot, there are a few things we can notice. Firstly, for smaller household sizes of six and below, there is a wide range of loan amounts, indicated by both the range of loan amount values and the interquartile range of the boxplots. This indicates high variability with several outliers, likely due to the fact that most homes bought fall within this category based on the histogram of household size, thus there is a large variety of homes in all price brackets that fall into these categories.

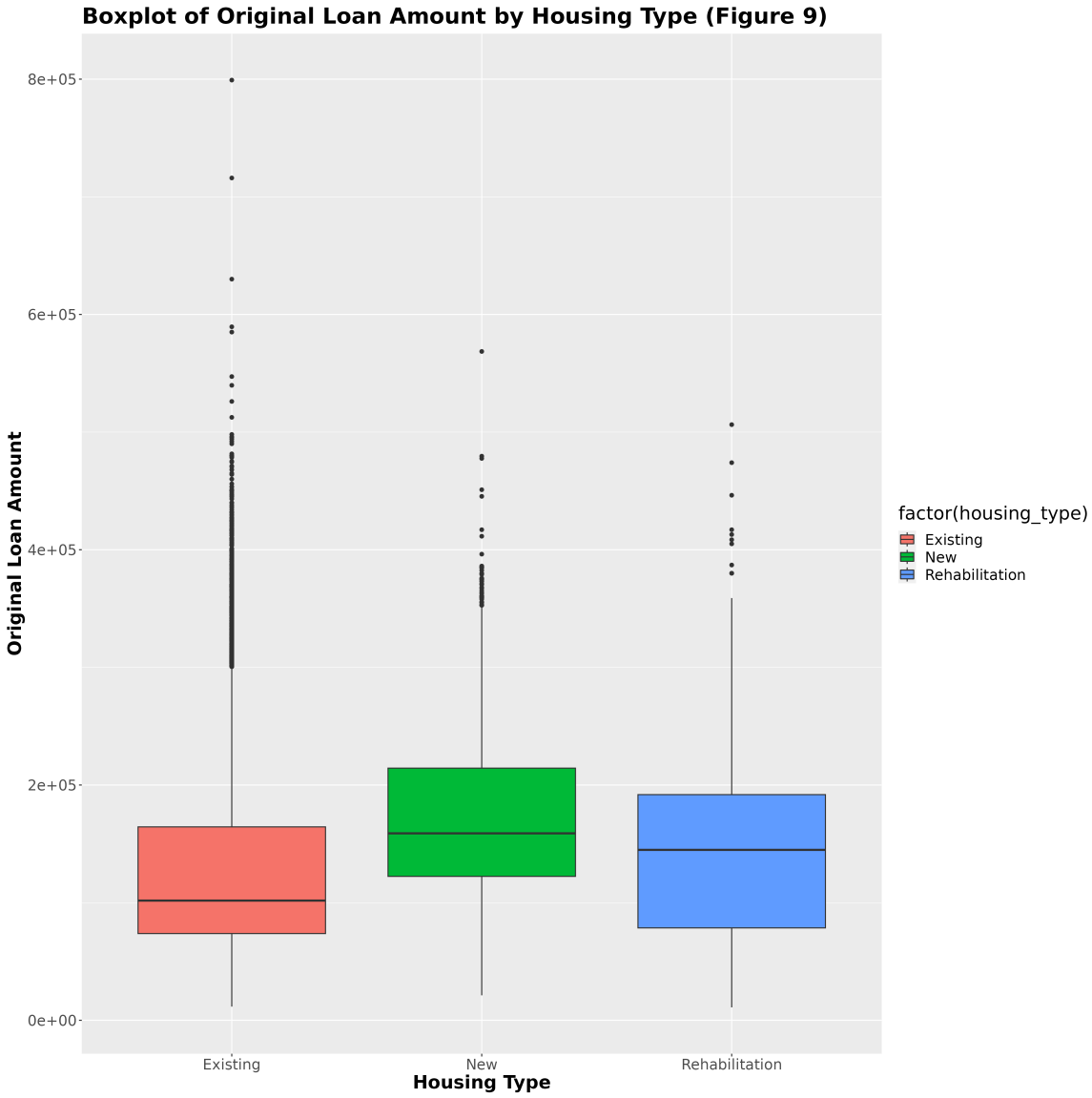
Additionally, we notice that the median loan amount for each household size is different and increases as household size increases. This indicates that there may be a positive relationship between household size and loan amount. This is supported by the fact that the middle 50% of the data appears to increase in loan amount as household size increases.



The majority of borrowers get a loan within 200,000 with majority get the loan around 100,000. It is reasonable because the majority of the clients of SONYMA are the low and moderate income families and also first-time home buyers, so it is often the clients' first time buying a significantly high-value property.



This boxplot illustrates that the loan amount varies wildly based on county. We see that most counties have a relatively small interquartile range, except for a few such as Bronx and Nassau that have colossal original loan amount ranges. This is likely due to the real-estate markets in each county, which may vary due to various geographical factors.



The boxplot indates that first-time low to moderate income people tend to get attracted by the new and rehabilitation properties more than the existing residences. It is reasonalbe because as the first-time buyer, people either tend to target the new houses if their financial background is not a serious problem (as only need to have a small loan) or if their financial background is not stable, they tend to select the rehabilitation with expected lower asking price.

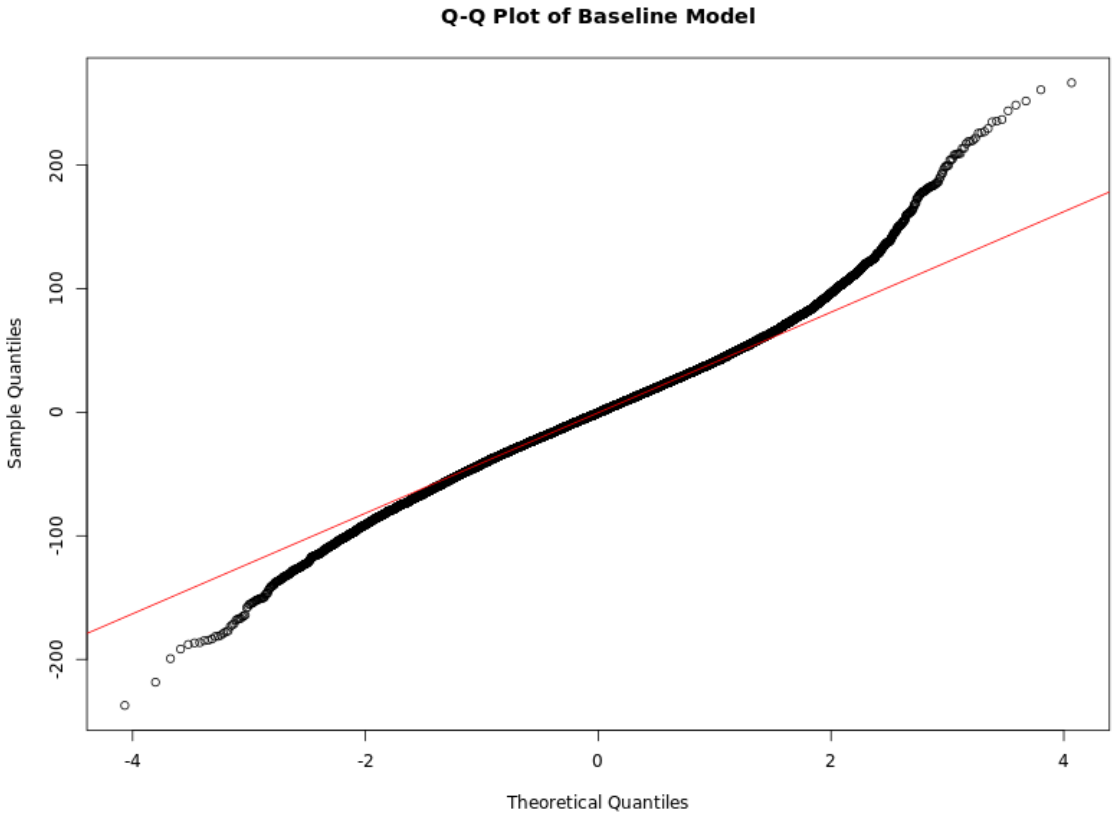
Model Selection

We utilized all forward, backward selection and stepwise selection as methods to choose the model and use adjusted R squared, Cp values as the decision criteria.

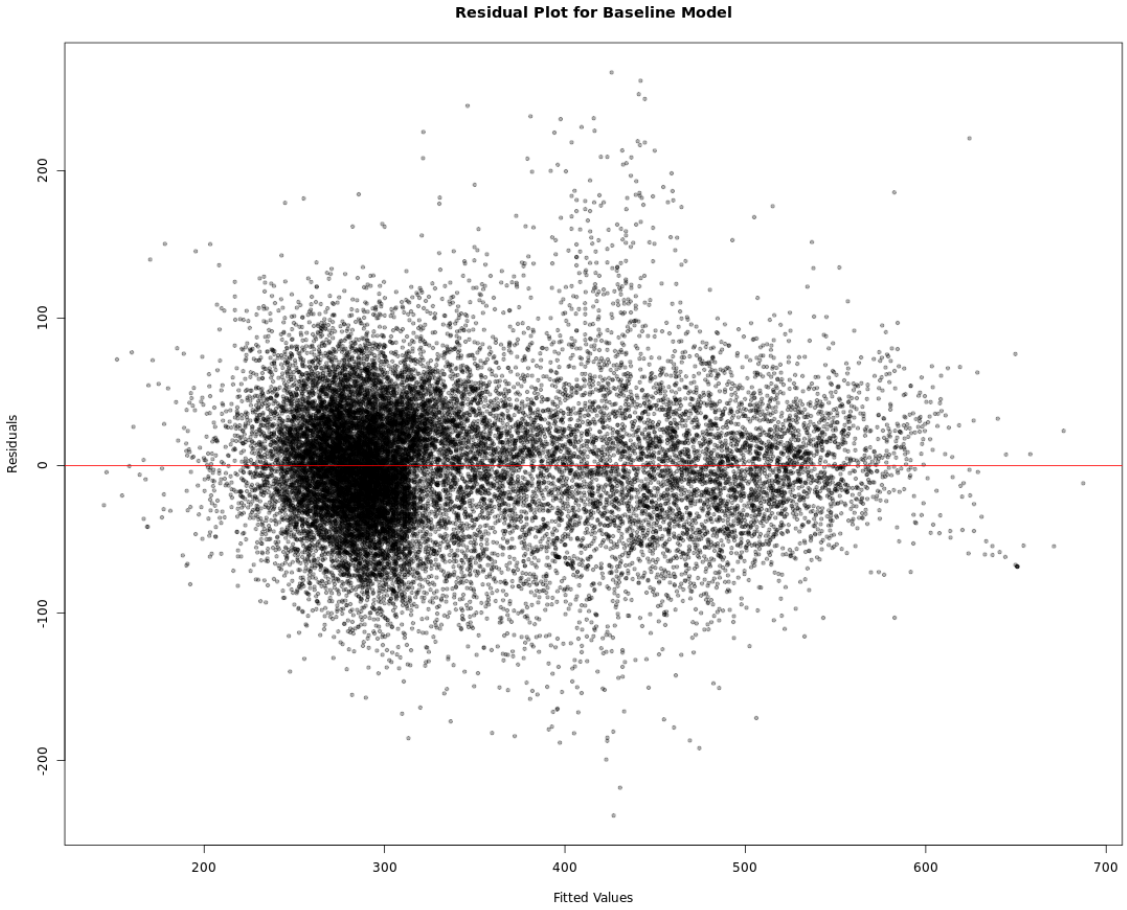
We first fitted a full model with no transformation as the baseline model for comparison and have an adjusted R squared value of 0.79 which is high for a baseline model. The next step is checking the assumptions of model through a QQplot and residual. There was no discernible pattern for the residual plot, but the qqplot shows the normality violation that the error distribution is right tail heavy, which is a problem.

Consequently, we had to apply a transformation to resolve this issue, and after examining feasible solutions such as using log for the responsible variables as well as the explanatory variables, we determined to choose to square root the response variable (original_loan_amount) and achieved an adjusted R squared of 0.78 which is lower and checking the assumptions of the model

Ultimately, all three methods suggest the same model as follows:



The new qqplot shows that the heavy tail still exist but now the error follows approximately normal distribution with some extreme values.



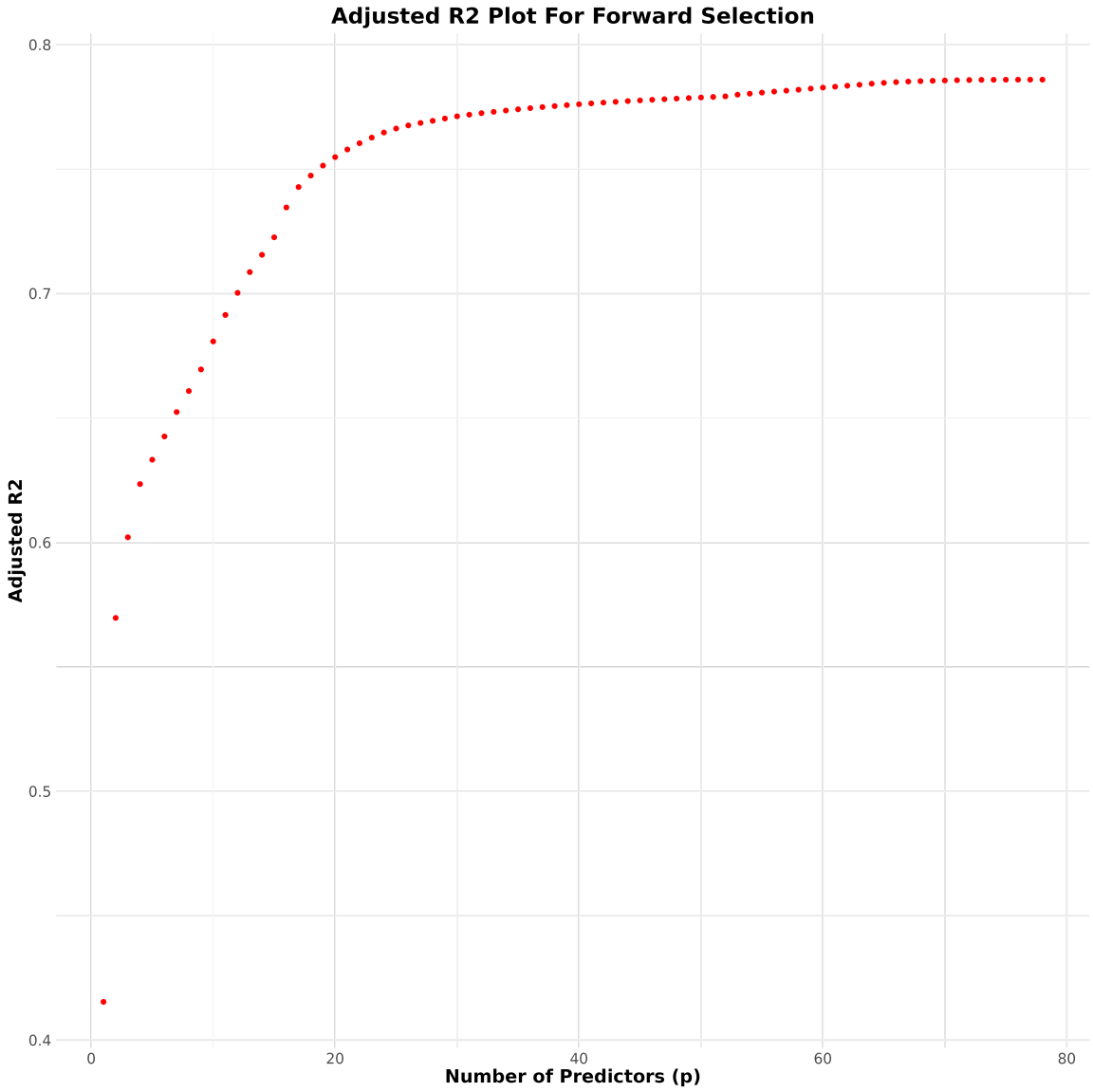
There is no discernible pattern in the residual plot, so we can say that this model did not violate any of the linear model assumptions.

Even though the adjusted R squared value is lower but not by a significant amount, the QQplot and residual showed that it is a more reliable model, so we decided to use the transformed model for both forward, backwards and stepwise selection.

Adjusted R squared criteria

Forward Selection

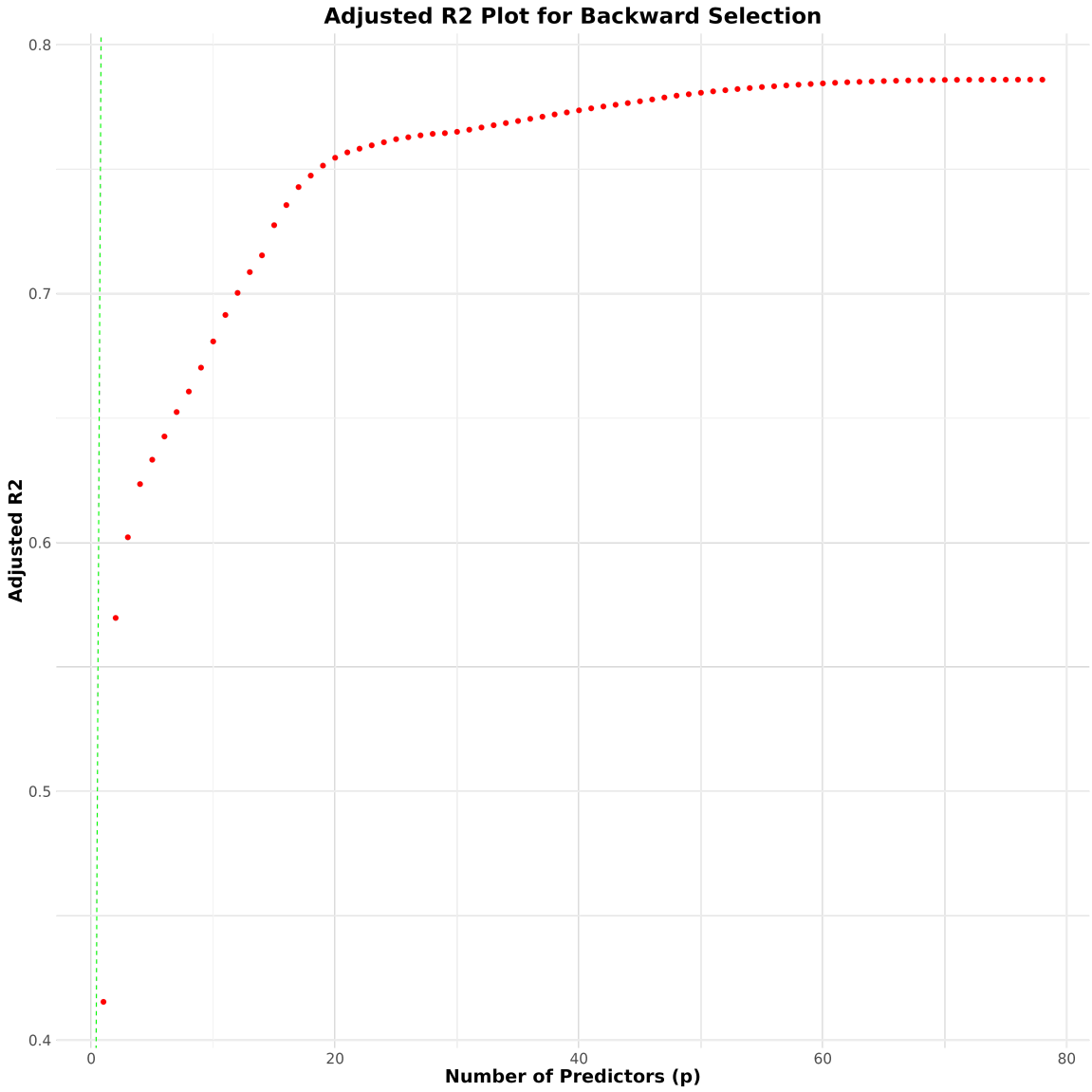
Now we can move to the forward selection method using the transformed model, we used regsubsets then plotted a gragh with x as the number of predictors and y as the adjusted R squared value



Using forward selection, we arrived at the revelation that the full model is the best performing model at the adjusted R square value of 0.78.

Backward selection

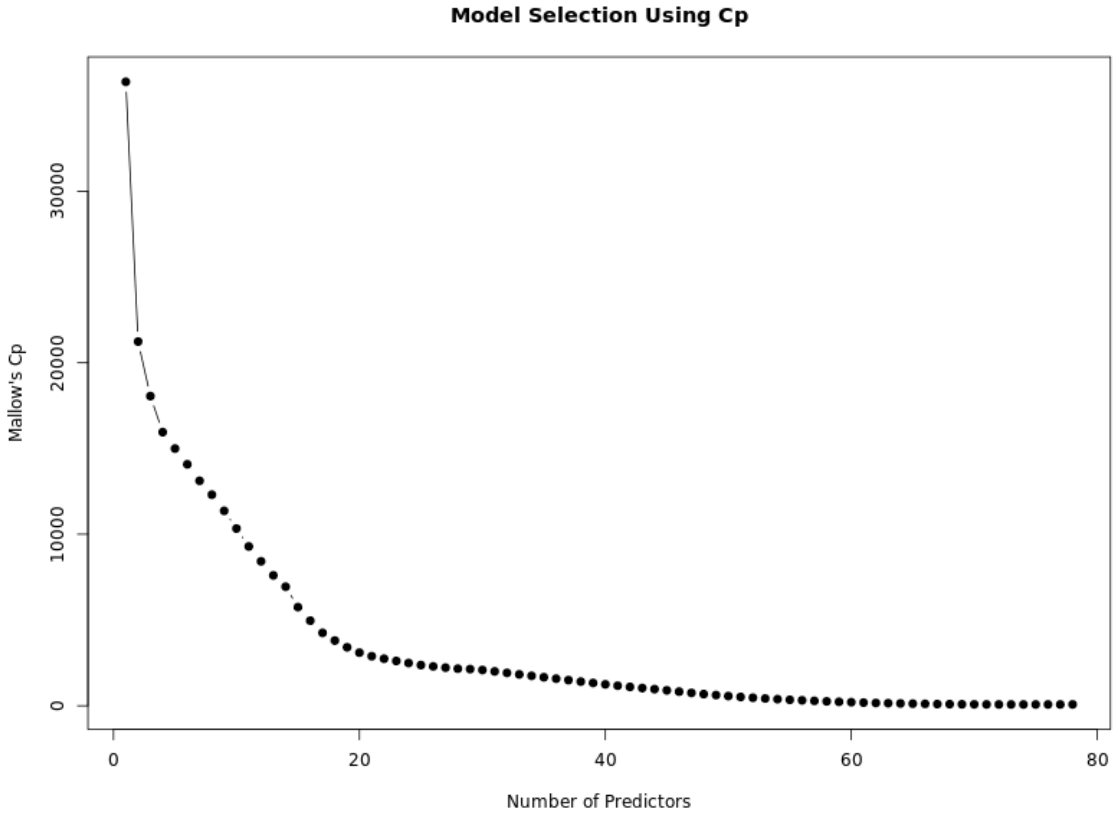
We move on to backward selection and also plotted the same graph



Backward selection arrived at the same conclusion as the forward selection method that the transformed full model is the best performing model at the adjusted R squared value of 0.78 which is also the same conclusion for the stepwise selection (in the code but the exact same values as the values as the backward and forward selection).

Mallows Cp criteria

Using Mallows Cp as the one of the criteria we also plotted the using number of predictors as the x and Cp value as the y:



The Cp gets lower and lower as the number of predictors increase and it get closest to the number of predictors with k = 79 and Cp = 79.

Final model decision

By the Cp value and adjusted R squared value, we can conclude that the full transformed model is the best performing model.

Conclusion

Our research investigates the relationship between geolocation and financial backgrounds and the mortgage loan amount that borrowers can get.

First, we dropped the extraneous columns and kept the admissible covariates for the response variable (original_loan_amount). Subsequently, we examine the most important covariate that affects the response variable is the SONYMA amount as well as other numeric covariates with a correlation heat map. We also utilized the boxplots to explain how the categorical covariates influence the response variable.

Then we tried all the backward, forward before checking with stepwise method (within the code modelling domain) to select the most appropriate model. The final model that we decided on was the transformed full model where we include all covariates and apply a square root transformation on the response variable (original_loan_amount)

$$
\begin{aligned}
\sqrt{\text{original\_loan\_amount}} = & -9372.4 \\
& + 4.744 \cdot \text{purchase\_year} \\
& + 1.504 \cdot \text{original\_loan\_to\_value} \\
& + 59.23 \cdot \text{loan\_type}_{\text{Step}} \\
& + 0.01085 \cdot \text{SONYMA\_amount} \\
& + 0.01971 \cdot \text{original\_term} \\
& + \sum_{\text{county}} \beta_{\text{county}} \cdot \text{county}_{\text{name}} \\
& + 6.552 \cdot \text{num\_of\_units}_{\text{2 Family}} \\
& + 37.31 \cdot \text{num\_of\_units}_{\text{3 Family}} \\
& + 44.58 \cdot \text{numof\_units}_{\text{4 Family}} \\
& - 29.14 \cdot \text{property\_type}_{\text{Condominium}} \\
& - 82.87 \cdot \text{property\_type}_{\text{Cooperative}} \\
& - 2.50 \cdot \text{property\_type}_{\text{Detached}} \\
& - 14.70 \cdot \text{property\_type}_{\text{Manufactured}} \\
& - 7.803 \cdot \text{property\_type}_{\text{Not Available}} \\
& - 11.96 \cdot \text{property\_type}_{\text{PUD}} \\
& - 3.084 \cdot \text{housing\_type}_{\text{New}} \\
& - 30.69 \cdot \text{housing\_type}_{\text{Rehabilitation}} \\
& + 2.747 \cdot \text{household\_size} \\
& + \varepsilon
\end{aligned}
$$

Interpretation of County Coefficients (Albany as Baseline)

In this linear regression model, **Albany County** is the reference category. The coefficients for other counties represent their effect on the **square root of the original loan amount**, compared to Albany, **while holding all other variables constant**.

Example interpretation:

If a property is located in **Kings** County, then — holding all other variables constant — the predicted square root of the original loan amount increases by **101.1** units compared to the **Albany** county.

Counties with Significantly *Higher* Loan Amounts (vs. Albany)

County	Coefficient	p-value	Interpretation
New York	141.3	<2e-16	Highest loan amounts
Nassau	118.6	<2e-16	Much higher loans
Queens	115.0	<2e-16	Much higher loans
Kings	101.5	<2e-16	Much higher loans
Richmond	96.4	<2e-16	Higher than Albany
Westchester	93.5	<2e-16	Higher than Albany
Rockland	91.1	<2e-16	Higher than Albany
Suffolk	88.5	<2e-16	Higher than Albany
Bronx	85.2	<2e-16	Higher than Albany
Putnam	62.9	<2e-16	Higher than Albany
Dutchess	49.5	<2e-16	Higher than Albany
Orange	44.8	<2e-16	Higher than Albany
Saratoga	9.1	0.0219	Slightly higher

Counties with Significantly *Lower* Loan Amounts (vs. Albany)

County	Coefficient	p-value	Interpretation
Allegany	-101.5	<2e-16	Much lower loans
Chautauqua	-100.1	<2e-16	Much lower loans
Chemung	-97.0	<2e-16	Much lower loans
Herkimer	-95.7	<2e-16	Much lower loans
Cattaraugus	-95.9	<2e-16	Much lower loans
Steuben	-84.7	<2e-16	Lower loans
Wyoming	-81.7	<2e-16	Lower loans
Oswego	-81.8	<2e-16	Lower loans
Lewis	-81.5	1.4e-05	Lower loans
Seneca	-81.7	<2e-16	Lower loans
Tioga	-78.2	<2e-16	Lower loans
Orleans	-78.6	<2e-16	Lower loans

...many more rural counties also show negative effects...

Counties with No Significant Difference from Albany

County	Coefficient	p-value	Interpretation
Columbia	+9.96	0.263	Not significant
Greene	-13.7	0.0635	Marginal
Hamilton	-50.7	0.0554	Marginal
Sullivan	-1.67	0.774	No difference

- **Urban & suburban counties (e.g., NYC, Long Island, Westchester)** tend to have **higher** loan amounts than Albany.
- **Rural/upstate counties** tend to have **lower** loan amounts.
- **Albany** serves as a **baseline/mid-range** county for comparison.

Limitation

The dataset contains columns of SONYMA operation (the source of mortgage, the time of acquiring the mortgage from the original lenders) and lacks explicit columns to explain the most important criterion of lending mortgage is clients' financial background (which is implicitly encoded in SONYMA) to infer the responsible variable which is the amount of loan mortgage money the borrowers would get. We have to assess the most import criterion through SONYMA amount (which is the amount of down payment assistance (DPAL) or closing cost assistance (CCAL) provided to the homebuyer as state in the data dictionaries) which is the result of assessment with internal IRS guidelines of the borrower credit reports, income information, bank statements, property appraisals, tax return, etc. Furthermore, this column contains 7516 Nulls value out of 28,528 data points that we had top drop all the nulls and have 1094 values = 0 which might not correctly represent all of the personal background of the borrowers.

Potential future research

For potential future research, if the dataset were to include explicit financial background data of the borrowers—such as income, credit scores, incomes, or savings, etc —then a more thorough investigation into mortgage risk, affordability, and the allocation of financial assistance could be conducted. The current reliance on implicit indicators (like SONYMA assistance amounts with many nulls) does not provide sufficient insight into the borrowers' actual financial conditions, limiting the accuracy and fairness of such analyses.