Project 2

Report

Government's Home Mortgage Data Analysis

Team 1:

Han Wu
Yunting Sun
Ruoyu Quan

Content

Part 1 Introduction

1.1 Background information

We obtained data from the FFIEC Home Mortgage disclosure Act (https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset), which was cleaned and briefly analyzed by members of Siete's analyst team. The original data included 1063316 rows and 101 columns of 2018 government's home mortgage data for New England and New York.

We will provide references and recommendations for promoting legislation to help first-time home buyers in New York and New England. To target at first time home buyers, we filtered the data that are potential first time home buyers, resulting in **361506 rows** of 2018 home mortgage data. Then we explore the subset to draw conclusions that point to any evidence of unfair practices or hurdles that first time home buyers encounter while trying to apply for a loan.

1.2 Filter analysis and explanation

Since we target at first time home buyers, for the purpose of accuracy, we have to narrow down the scope of analysis within data based on the assumptions we made of the first time home buyer's characteristics. The following columns were filtered as they are not applicable for the target dataset:

1. **Business_or_commercial_purpose**
   Values of '1' under this column were filtered out because these represent loan applications that are primarily for a business or commercial purpose. For first time home buyers, the main reason for purchasing was the desire to own a home of their own.
2. **Occupancy_type**
   Values of '1' under this column were retained because home buyers usually buy their first home to serve as their principal residence.
3. **Loan_purpose**
   Values of '1' under this column were retained because these represent loan applications made for the purpose of home purchase.
4. **Action_taken**
   Values of '4' and '5' were filtered out because these represent incomplete transactions. We will not focus on the withdraws and incomplete applications.

Based on our data on hand, we believe these four filters give us a definitive subset that can be used as a sample for first time home buyers in New England and New York.

Limitations:

Besides, we have also considered some other variables that can be experimented with which are age, income and property value. Generally, first time home buyers are younger, having low income, and buying less expensive houses than move-up buyers**.** However, we do not have sufficient information to make an informed decision to filter first time home buyers based on age, income and property value.

Age and income are tricky factors to judge first time home buyers. We can't assume all the young people, or people with low income are the only group of first time home buyers. For some people having low income, they have to purchase property in a late time during their life time due to financial reason. In this situation, it will take them a longer time to have the down payment.

On the other side, people with high income could also be the first time home buyers. Some people are highly possible to have a high income level due to the working field which provides them higher average income than others, like people working as information technology and lawyers.

Accordingly, property value is also another factor we evaluated in filtering process. Due to the situation we discussed above, people having high income in a promising field are highly possible to purchase high value property. This is violated our assumption at the beginning that first time home buyers purchase low value property.

Unfortunately, we don't have such information on applicants background. If we only keep people having low income, age under 35, and purchasing low value of property, this would give us a misleading subset which is highly possible to exclude people who are first time buyers in other perspectives.

Therefore, we only use the columns we listed and explained above to filter first time home buyers based on the value and definition of dataset.

Part 2 Statistical Analysis

2.1 Basic dataset summary

Before we start statistical analysis of first time home buyers data, we screen on all the variables including in the dataset to confirm relevant factors to be analyzed.

2.1.1 Columns selection for analysis

The original home mortgage data includes the following main information:

1. **Applicants information:** Age, Sex, Race, Ethnicity, Location, Income, action taken
2. **Loan information:** Property Value, Loan Amount, Loan Term, Interest Rate, Loan Type, Total Loan Costs, debt-to-income ratio, property to income ratio
3. **Covered loan characteristics:** Lien status, Reverse mortgage, Open end line of credit
4. **Payment method:** Negative amortization, Interest-only payments, Balloon payment
5. **Others:** Denial reason**,** Pre approval, conforming loan limit

The first finding is that all of the first time home applicants have a common pattern in loan characteristics, payment method, and others:

- Almost all covered loans are secured **by a first lien**, **not a reverse mortgage** and **not an open end line of credit.**
- Almost all contractual terms include **no negative amortization**, **no interest-only payments and no balloon payment**.
- Almost all the applications are **not getting pre approvals and not beyond the range of conforming loan limit.**

Therefore, we don’t need these columns to perform any further analysis. What we will focus is to analyze the stories behind loan amount, property value, income, debt to income ratio, property to income ratio, total loan costs, denial reasons and other information to make recommendations to help first time home buyers.

2.1.2 Overview of application

First, we take a quick look on the numerical summary of loan information to get a rough understanding of loan and property situation that our applicants face:

Figure1. Numerical summary of loan information

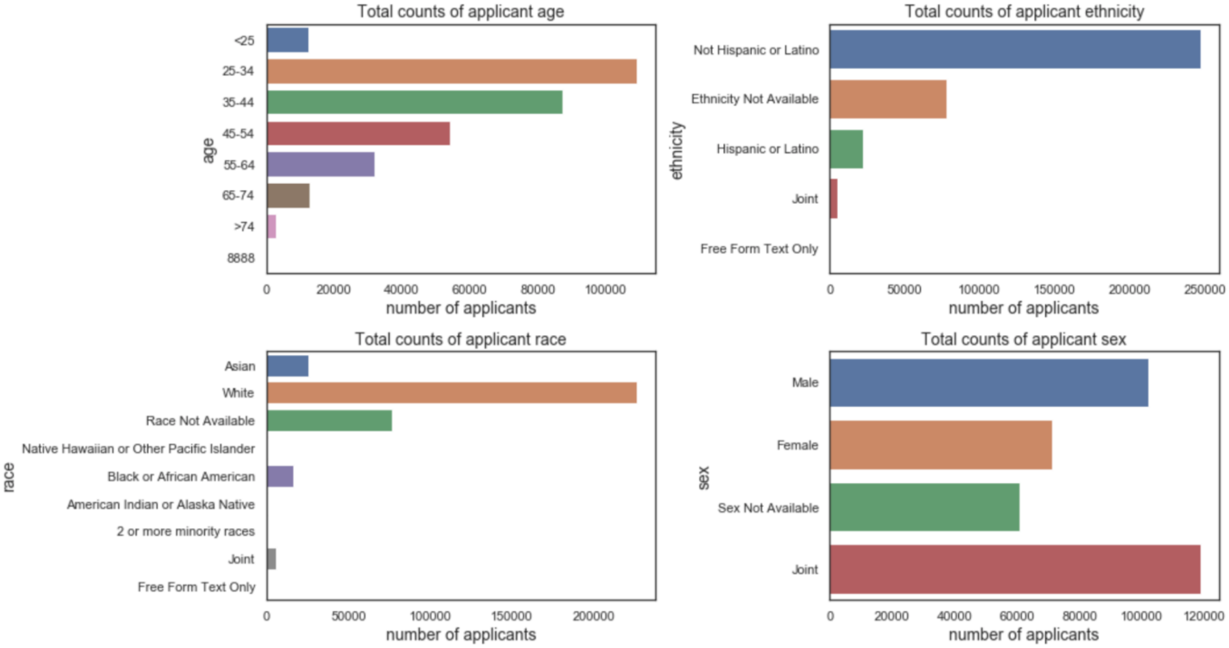
Mean Income(in thousands)	**$124.57**
Mean Property Value(in thousands)	**$416.55**
Mean Loan Amount(in thousands)	**$324.46**
Mean Total Loan Costs(in dollars)	**$5532**
Mean Loan Term	**347 months**
Mean Interest Rate	**4.59%**

Figure 2. Percentage summary table of loan type

Loan Type	**Percentage**
Conventional (not insured or guaranteed by FHA, VA, RHS, or FSA)	**73.1%**
Federal Housing Administration insured (FHA)	**20.2%**
Veterans Affairs guaranteed (VA)	**5.4%**
USDA Rural Housing Service or Farm Service Agency guaranteed (RHS or FSA)	**1.3%**

Then, we summarize the geographical information to know who they are:

Figure 3. Bar charts of geographical information of applicants



Then, we know that the people aged 25-34 are the most high proportion of groups to apply for loan for their first property, followed by people aged 35-44, 55-64. People whose age below 25 and over 74 are the minority to purchase property for the first time. Seldomly, people aged over 74 would purchase property for the first time.

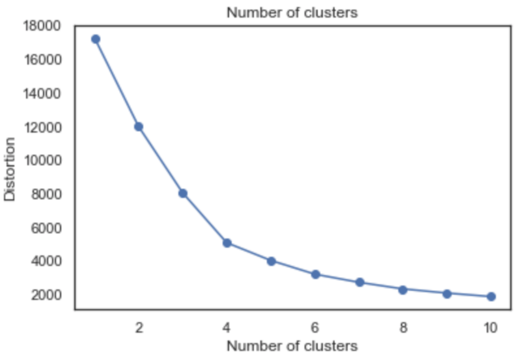
We can also easily find the ethnicity distribution that most people are not Hispanic or Latino and there is only a very small portion of applicants are Hispanic or Latino or joint. For the race distribution, most applicants are the White, then Asian and Black or African American. The minorities only take a very little part and are hard to show on the graph, including people of some Islander, 2 or more minority races or Joint races. Lastly, we find Joint gender applications are most common to see, then followed by Male and Female.

2.2 Clustering analysis of loan behavior

This part explores loan behaviors of first time home buyers defined by previous filtering conditions assumed. To start, K-means clustering technique is used to cluster mortgage applicants in terms of property value, loan amount, income and whether an application got approved. Then, we summarize trends and features among different age and race groups based on our clustering.

2.2.1 Clustering method of k-means

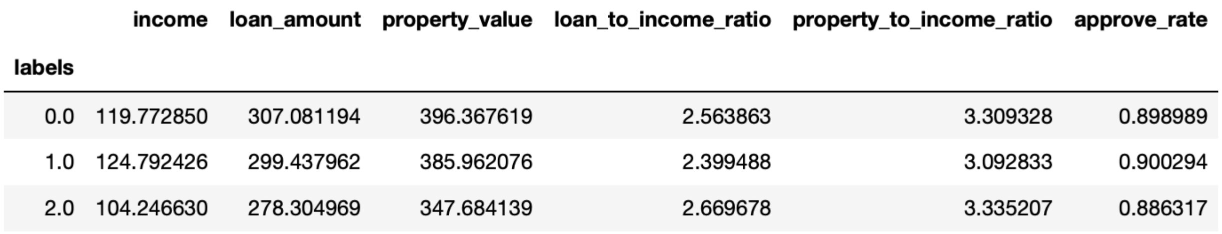
Figure 4. Number of clusters by elbow method



We use the elbow method to confirm the number of clusters. The elbow method is a useful graphical tool to estimate the optimal number of clusters. The idea behind the elbow method is to define the value of k where the distortion begins to decrease rapidly.

As it was shown on the line graph of 'Number of clusters', the number of clusters should be 4. However, we see the result of clustered groups by 4 and find that there are 2 groups that are too close, illustrating a not significant difference. Therefore, we decide to take the optimal number of clusters as 3.

Figure 5. Classification of applicants by k-means method



	income	loan_amount	property_value	loan_to_income_ratio	property_to_income_ratio	approve_rate
labels						
0.0	119.772850	307.081194	396.367619	2.563863	3.309328	0.898989
1.0	124.792426	299.437962	385.962076	2.399488	3.092833	0.900294
2.0	104.246630	278.304969	347.684139	2.669678	3.335207	0.886317

Then, we classify our data into 3 different groups to explore the behaviors of first time home buyer applicants. The following pivot table shows average income level, house value, as well as their ratios, according with approval rate in each group:

The general features of three groups are:

- Group 1 - medium income, high approval rate, high debt ratio
- Group 2 - high income, high approval rate, lowest debt ratio
- Group 3 - low income, low approval rate ,high debt ratio

Correspondingly, Group 3 has the lowest approval rate. Next, we want to find features of loan behaviors in different ages and race for these groups.

2.2.2 Clustering analysis by age

Here are subplots of income and property value in each group by age:

Figure 6. Boxplots of income distribution by age

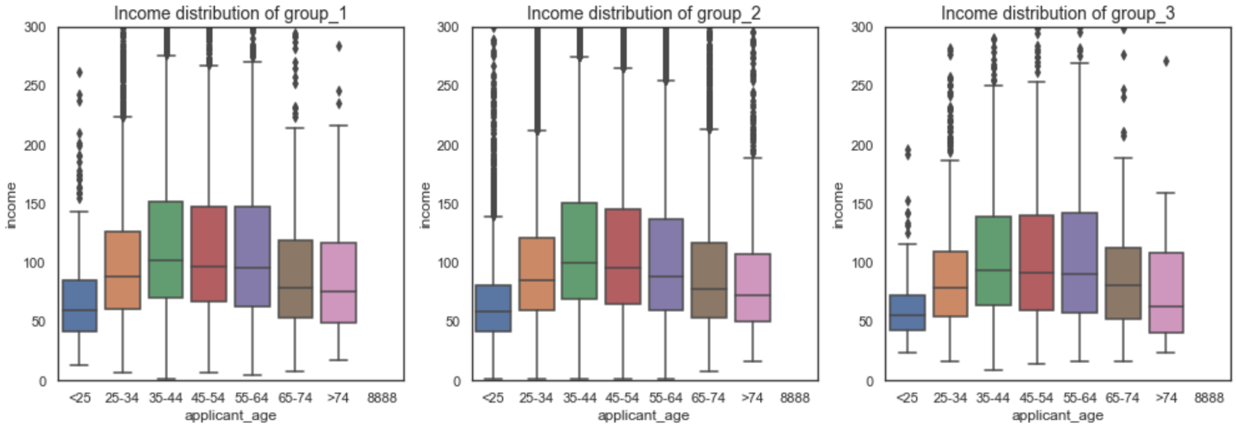
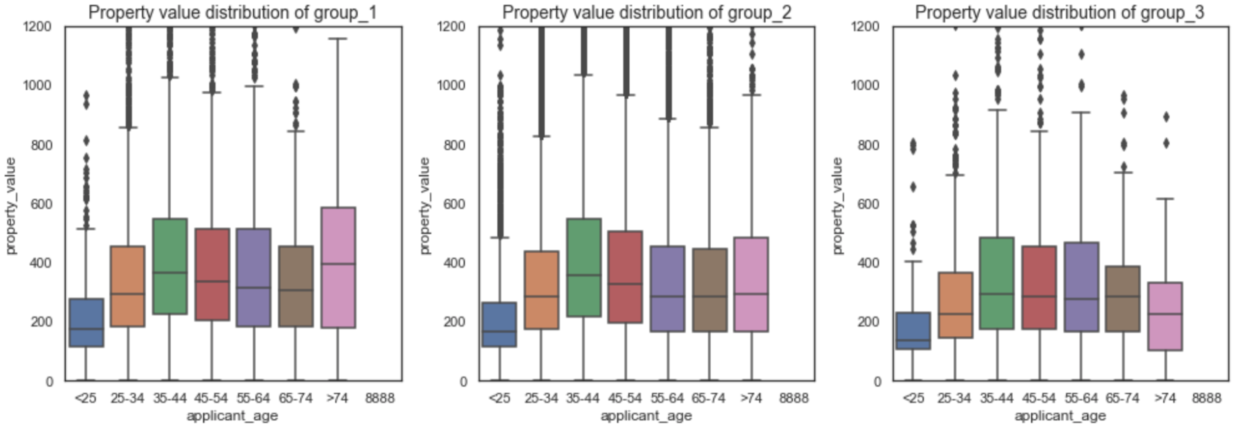


Figure 7. Boxplots of property value distribution by age



The income distribution trends are similar in all groups. As age goes up, income increases, and gets steady at the middle age of 35-44, 45-54, and 55-64, then declines after 65. Group 2 shows a strong trend in people aged 35-44 reaching an income peak. So, **middle aged people, having relatively higher income are the majority house buyers as the box plot shows, especially for the rich.**

What comes to a difference is the property value. It is clear to see people whose age above 74 , having medium and high income tends to buy high value property. But this is not the case with people whose age above 74, having low income. It could be understood that the **elder and rich are more likely to buy high value property than others. They own a well financed situation at their age. In contrast, for elder and poor people, they can’t afford to buy high value property.**

2.2.3 Clustering analysis by race

Here are subplots of income and property value in each group by race:

Figure 8. Box plots of income distribution by race

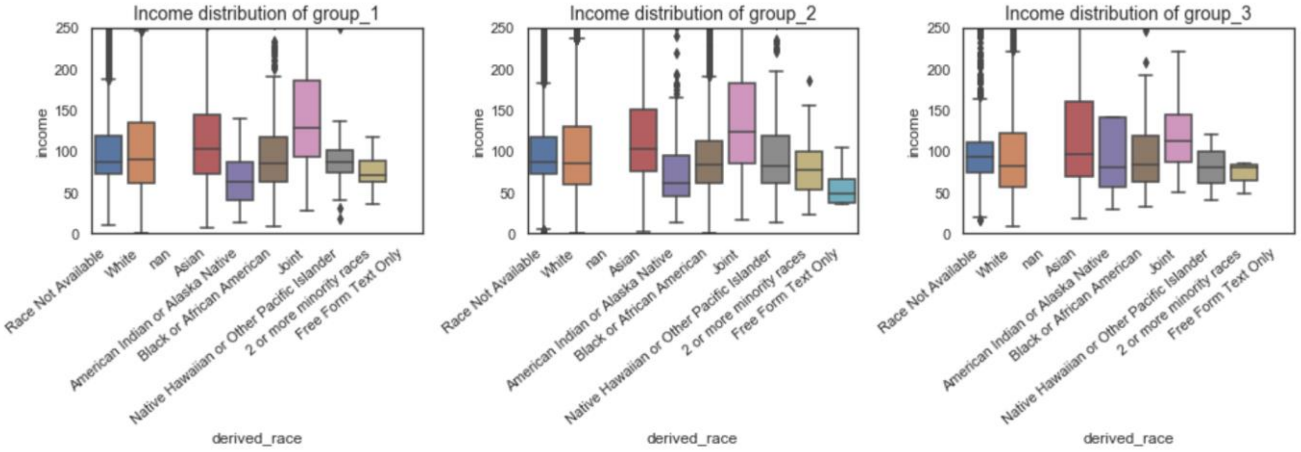
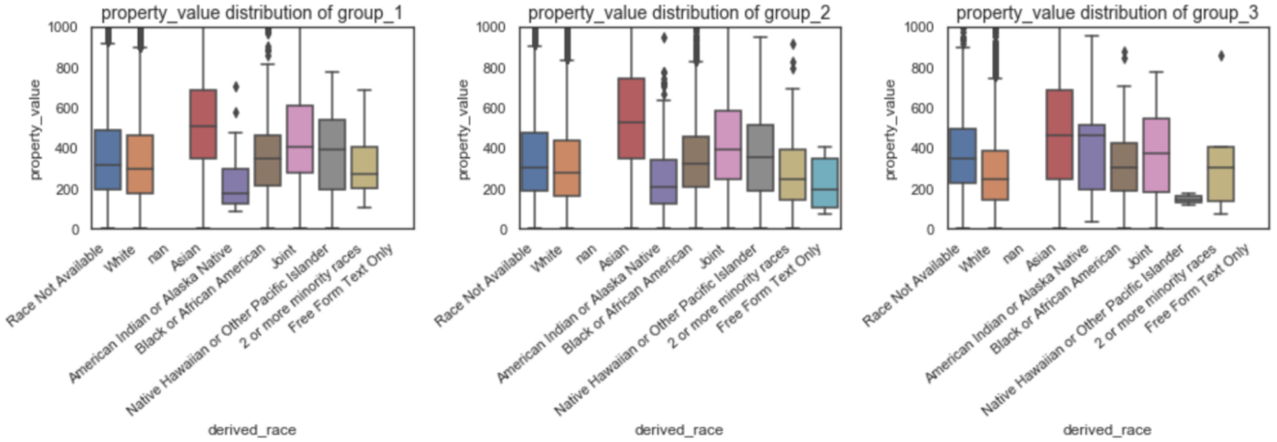


Figure 9. Box plots of property value distribution by race



We can find the same trend of income for both group 1 and group 2, who are from relatively high income groups, so as to property value. For these groups, the majority are applicants of joint races, then coming with Asian and White. But, this trend ends up in group 3, low income group.

Foe general trend, joint race applicants having highest income but not buying highest value of property. **Asian buy the highest value of property and their income is also relatively high in all groups.** But for some people in group 3, **joint race applicants are more conservative.**

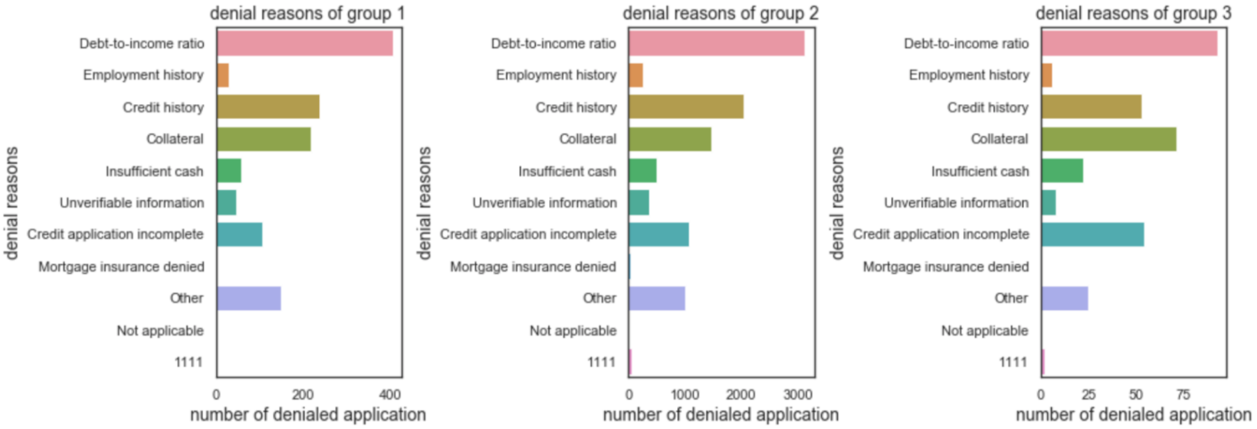
In all races from different levels of income, their buying value is generally based on income level. But one exception is low income **applicants from Native Hawaiian or other Pacific Islander.** If their income is under 100,000, the property value for them is sharply decreasing to under 200,000.

2.3 Denied application analysis

In order to make recommendations to legislators to help first time home buyers, the following analysis is to focus on denied applications by comparing denial reasons among clustering groups. We aim to confirm detailed information about applications failure to get loans. Then use a pivot summary table to make conclusions by age and race.

2.3.1 Denial reason overview

Figure 10. Bar charts about denial reason in each group



The bar plots of denial reasons look the same for both group 1 and group 2, with ‘debt-to-income ratio as first, credit history as second and collateral as third. All the other reasons took a small portion. However, for applicants from low income level groups, they were rejected for ‘debt-to-income ratio’, followed by ‘Collateral’, then ‘Credit history’. Besides, they also have one more problem, which is ‘credit application incomplete’. This information suggests 2 aspects:

1. **low income applicants don’t have enough high value assets as collateral.**
2. **low income applicants are not knowledgeable with the process of credit application.**

Thus, we suggest legislators to build up a program for educating low income first time home buyers on credit evaluation, preparation and application. The way of educating could be text information, or online video link sending by email when applicants create an application account.

Moreover, due to low income people have relatively less assets to be collateral. Legislators could find ways to figure out the way to be collateral other than assets. For example, legislation could require low income applicants to use financial products of deposits to be collateral, or to make a reverse mortgage to give away the right of property after an amount of years.

2.3.2 Analysis of debt to income ratio

From the denial reason bar charts, we got ‘debt-to-income ratio’ as a major and the most common reason for all failed applications no matter how applicants are divided. It is common sense that applicants must be evaluated for the ability to repay their loans. If the debt to income ratio is currently high, which means it is possible that the future income would not cover loan amounts so that this would affect the success of application. **Thus, we aim to define what factors could lead applicants to have a high debt to income ratio. Then, we specify these buyers with low probability to be approved in an age-level and race-level, respectively**. By understanding the picture behind this, legislators could find ways to help them on mortgage.

A. Simple linear regression analysis of loan amount

First, the plot on the left side is the scatter plot of the loan amount and property value showing the distribution of all the paired points. An upward trend of distribution of points suggests a linear relationship between loan amount and property value. Next step, we plot the linear regression plot as shown on the right side.

Figure 11. Scatter plot of loan and property

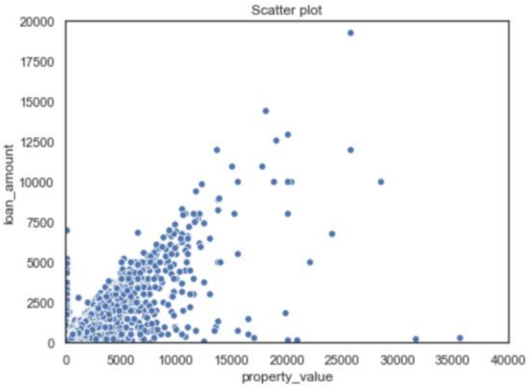
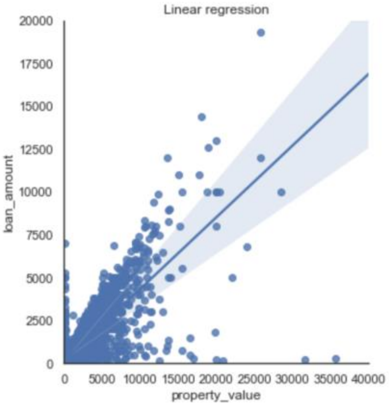


Figure 12. Linear regression of loan and property



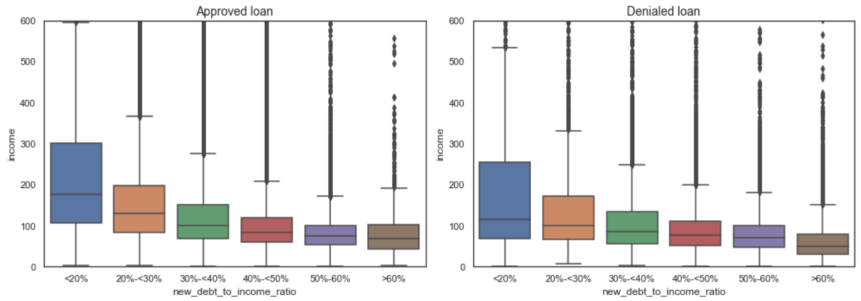
As the Linear regression graph shows, an upward trend line exists between loan amount and property value, indicating a positive relationship between these two variables. Therefore, we get a general rule about the factor affecting loan amount. We can conclude that **there is evidence showing as the property value increases, people would have a higher loan amount.**

But we also find some outliers at our Linear regression and Scatter plot. Therefore, in the later part, we would **recognize and understand applicants who would not behave as what we predict** and then give suggestions to legislators to improve the experience on loan mortgages.

B. Comparison of approved and rejected loan on income and property value

Except for property value, we assume that there are some other factors that would affect the value of the loan amount, for example, **income**. Income is as important as property value influencing the debt-to-income ratio. Here, generally we compared income and property value of approval and denial groups at various level of debt-to-income ratio:

Figure 13 . Comparison of income against debt to income ratio between approval and denial

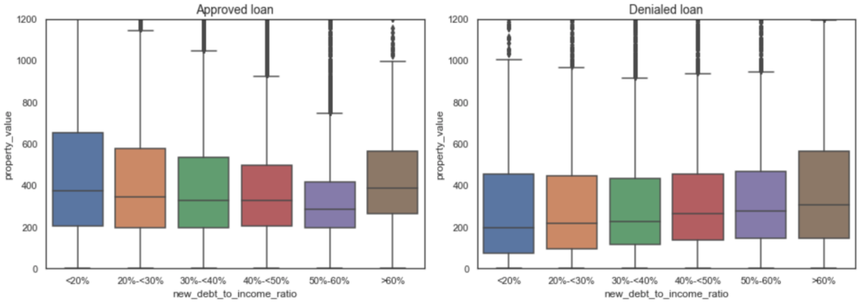


Firstly, this boxplot on the left implies that as the debt-to-income ratio increases, the income declines obviously no matter whether the loan is approved or not.

This suggests that even applicants who have a relatively low income and higher debt-to-income ratio, they have the opportunity to get approval on their loans. Similarly, for high income applicants, they also probably got denied on loan application even though their debt ratio is low and income is high. So, **income seems not a crucial factor affecting the result of approval**. **Then, what indeed does determine the result?**

Then, we plot the comparison between approved and denied groups on property value distribution in each debt-to-income level:

Figure 14 . Comparison of property value against debt to income ratio between approval and denial



a. We see that in the approved loan group, as the debt-to-income ratio increases, which suggests the risk in repaying loans, the property value gets smaller sharply until 50%-60%. This implies **the lenders would prefer to give loans on high property value for people with low debt-to-income ratio. Even for applicants whose debt ratio is high, they still have a chance to be approved on relatively low value of property.**

b. If we look at the denied group, a different picture shows up. They purchase high value property at a high debt-to-income ratio. And as the ratio gets a slight increase from 30%, the property value also goes up. This suggests that for these group people:

1.They need to buy a property no matter what the financial situation they have.
2.They don't have a clear picture about how much they should spend on property which is proper for them financially justified.

Here, we recommend that legislation could require all the applicants to complete a preapproval before originating an application. Currently, preapproval is not a must have step in loan mortgage application. But, it can help first time home buyers understand the current situation they stand to make a formal decision. This way could also help them save time and money on unnecessary failures on loan application due to the loan costs.

We also believe that first time home buyers have demand on houses for families. Some low income applicants should get a proper guidance on screening property. Alternatively, the group of people who might have a high risk to be rejected on loan application should get some extra program on loan issues, like interest rate policy and payment method policy.

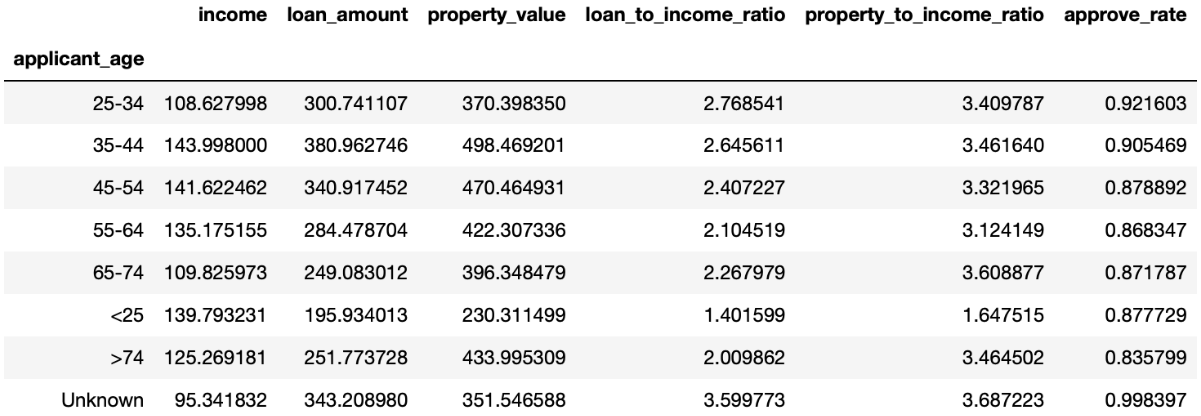
Next, we would specify what this group people look like to help legislators find them and give a hand to help them.

C. Approve rate in age group
we have discussed in previous chapter about the loan behaviors in different age and race groups, we have concluded that:

1. **middle aged people, having relatively higher income are the majority house buyers as the box plot shows, especially for the rich.**
2. **Interestingly, the elder and rich are more likely to buy high value property than others. They own a good finance situation at their age. In contrast, for the elderly and poor, they can't afford to buy high value property.**

This finding seems to point out where the problem exists: the elder group has the demand to purchase a house at their age no matter their income level. As we discovered earlier, the elder having relatively high average income prefer to buy high value property more than what the middle aged people having higher income buy. What's more, the elder having low average income tend to purchase low value property, suggesting they need their own property to stay. At this moment, let's see how is the approval rate for people in different age level as the following summary table shows:

Figure 15. Summary table of mortgage ratio with approve rate by age



applicant_age	income	loan_amount	property_value	loan_to_income_ratio	property_to_income_ratio	approve_rate
25-34	108.627998	300.741107	370.398350	2.768541	3.409787	0.921603
35-44	143.998000	380.962746	498.469201	2.645611	3.461640	0.905469
45-54	141.622462	340.917452	470.464931	2.407227	3.321965	0.878892
55-64	135.175155	284.478704	422.307336	2.104519	3.124149	0.868347
65-74	109.825973	249.083012	396.348479	2.267979	3.608877	0.871787
<25	139.793231	195.934013	230.311499	1.401599	1.647515	0.877729
>74	125.269181	251.773728	433.995309	2.009862	3.464502	0.835799
Unknown	95.341832	343.208980	351.546588	3.599773	3.687223	0.998397

We can find people over the age of 74 have the lowest approval rate on loan application. From the table, we conclude that:

1.the average income level for applicants over the age of 74 is the lowest, about $125,000.
2.However, their property value is relatively high, which is $434,000. We know from previous findings that some rich elder people have higher income to buy high value property. Those with low income would buy low value property.
3. The loan to income ratio is relatively low compared to other age groups. This suggests that older people have a deposit for them to buy whatever value property.

Thus, we can say those elder people over 74 can be grouped into 2 categories: rich with lots of deposit to purchase a house and poor with less money but can buy a low property value house with a small amount of loan. It looks like they are rational in buying property and have preparation on this issue. However, they can’t get a supportive loan as easily as other age groups. Maybe, it is reasonable for lenders to consider that they don’t have time as long as younger people to work and have steady income so that it will affect the application result.

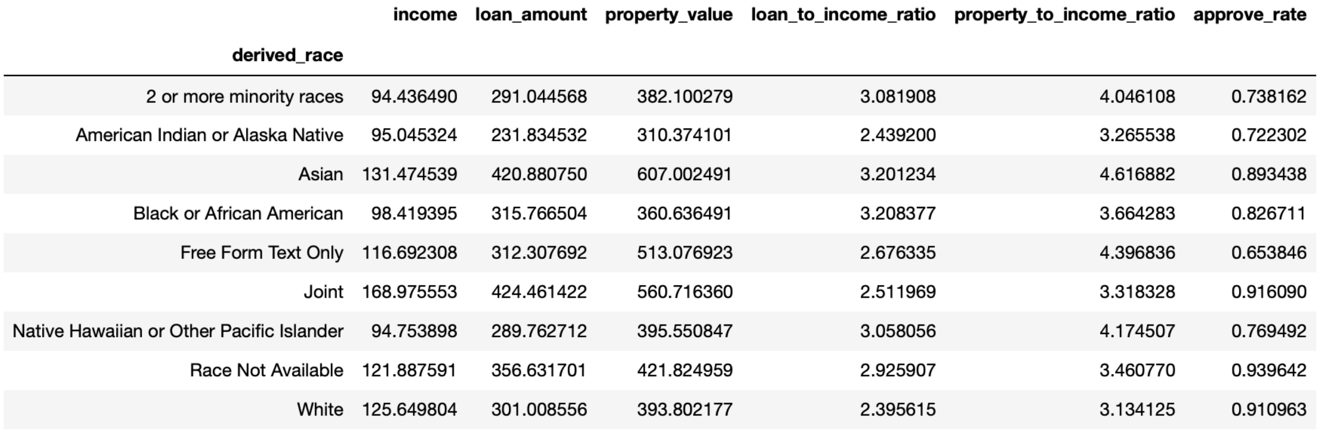
From this perspective, we recommend legislation to help this group of people to get a justified consideration on loan application for property in a more specific way on interest rate, collectoral, reverse mortgage, and any other payment method. As what we discovered from our dataset, the payment ways for all applicants are the same. The results show that they didn’t use negative mortgage, interest rate payment or balloon payment. And the typical loan term for all loan applications is the same as 30 years around.

Legislator could make a more flexible loan method for this group to change some policies discussed above. One example is to use a reverse mortgage which can reduce the risk of being unable to repay the loan.

D. Approve rate in race group

The following summary table list the approve rates with critical values about mortgage in different race group:

Figure 16. Summary table of mortgage ratio with approve rate by race



	income	loan_amount	property_value	loan_to_income_ratio	property_to_income_ratio	approve_rate
derived_race						
2 or more minority races	94.436490	291.044568	382.100279	3.081908	4.046108	0.738162
American Indian or Alaska Native	95.045324	231.834532	310.374101	2.439200	3.265538	0.722302
Asian	131.474539	420.880750	607.002491	3.201234	4.616882	0.893438
Black or African American	98.419395	315.766504	360.636491	3.208377	3.664283	0.826711
Free Form Text Only	116.692308	312.307692	513.076923	2.676335	4.396836	0.653846
Joint	168.975553	424.461422	560.716360	2.511969	3.318328	0.916090
Native Hawaiian or Other Pacific Islander	94.753898	289.762712	395.550847	3.058056	4.174507	0.769492
Race Not Available	121.887591	356.631701	421.824959	2.925907	3.460770	0.939642
White	125.649804	301.008556	393.802177	2.395615	3.134125	0.910963

From the income column, we can find the minorities are all having low income levels, including ‘2 or more minority races’, ‘Native Hawaiian or Other Pacific Islander’, ‘American Indian or Alaska Native’, ‘Black or African American’, along with low approval rate.

Obviously, for both of 2 or more minority races and Islander, they also have a very high property to income ratio and loan to income ratio even though their average property value is not high. **This implies that their income level is too low for the current property market.**

But no one can eliminate their demand for property. Therefore, we recommend legislators to help these minorities to help them find the property based on their income level and financial situation. One blueprint is to build a community with low construction and land cost only for these minor race people to live in. The trade off might be a long travelling distance to downtown or city. Thus, this should be a social problem for the government to find a way to provide a basic place for minorities to set their homeland.

From the other perspective, legislators should also encourage relevant government departments to improve the overall education level of minorities for the next generations to increase the average level of income. Thus, it could solve the problem at the beginning.

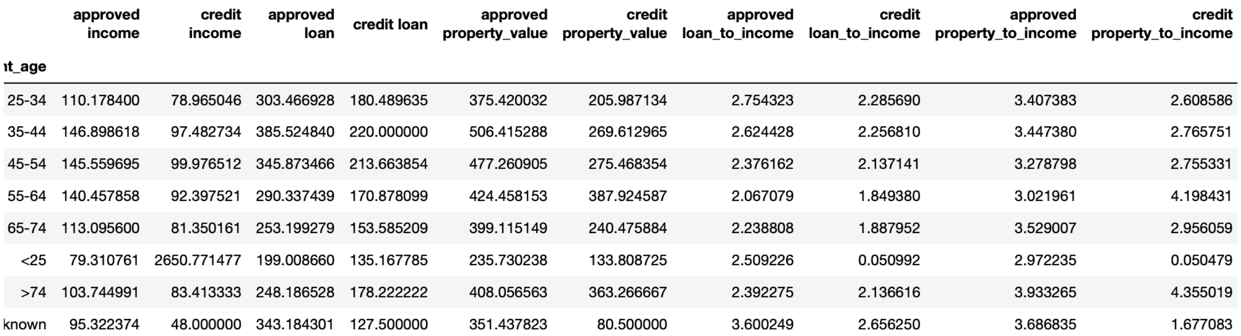
2.3.3 Analysis of credit history

We have limited data information of credit score about all the applicants. The only relevant data is the value of denial reason as ‘credit history’. Thus, we make a comparison between approved loan applications with denied loan applications for credit history by age-level and race-level.

A. Comparison of approval and denied by credit by age-level

The following pivot table summarize the comparison in different age groups:

Figure 17. Summary table of comparison between approval and denied for credit by age



	approved income	credit income	approved loan	credit loan	approved property_value	credit property_value	approved loan_to_income	credit loan_to_income	approved property_to_income	credit property_to_income
ıt_age										
25-34	110.178400	78.965046	303.466928	180.489635	375.420032	205.987134	2.754323	2.285690	3.407383	2.608586
35-44	146.898618	97.482734	385.524840	220.000000	506.415288	269.612965	2.624428	2.256810	3.447380	2.765751
45-54	145.559695	99.976512	345.873466	213.663854	477.260905	275.468354	2.376162	2.137141	3.278798	2.755331
55-64	140.457858	92.397521	290.337439	170.878099	424.458153	387.924587	2.067079	1.849380	3.021961	4.198431
65-74	113.095600	81.350161	253.199279	153.585209	399.115149	240.475884	2.238808	1.887952	3.529007	2.956059
<25	79.310761	2650.771477	199.008660	135.167785	235.730238	133.808725	2.509226	0.050992	2.972235	0.050479
>74	103.744991	83.413333	248.186528	178.222222	408.056563	363.266667	2.392275	2.136616	3.933265	4.355019
known	95.322374	48.000000	343.184301	127.500000	351.437823	80.500000	3.600249	2.656250	3.686835	1.677083

It is easily to find that:

1. People who have credit issues are those having low income no matter what age they are.
2.The property value and loan amount for these people are relatively much lower than those approved applicants. The corresponding ratios are not high.
3.However, people between 55-64 and over 74 have a high property to income ratio but a normal loan to income ratio, as we find they purchase high value of property but having low income.

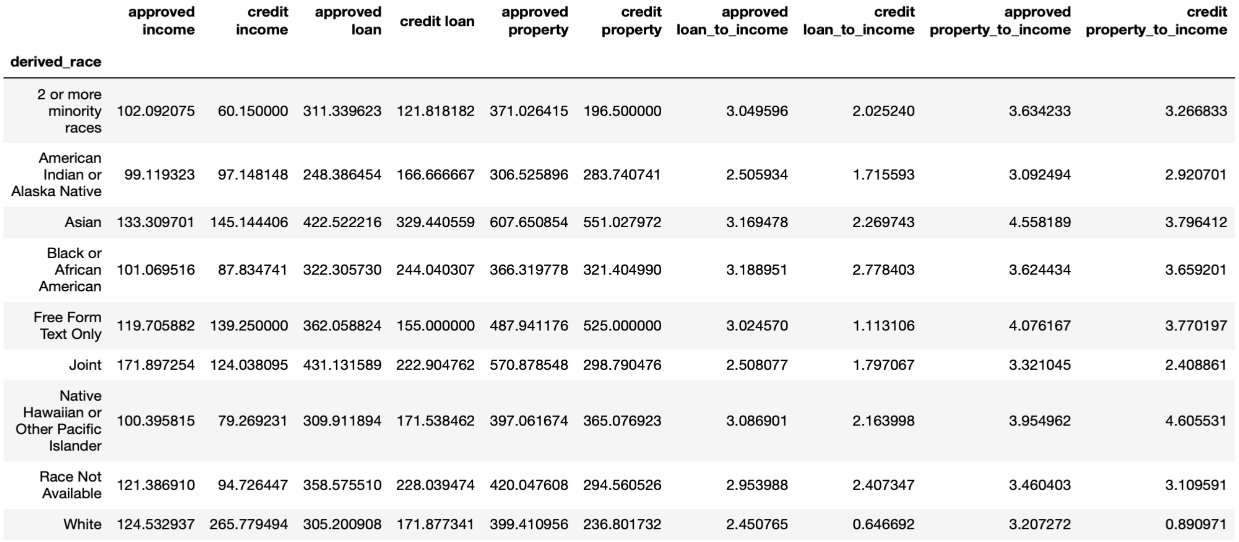
The third findings suggest that people having credit issues and whose age are between 55 to 64 and over 74 may have a big amount of cash on hand and borrow less **due to their loan to income ratio being low and property to income ratio being high**. Do they have enough deposit to pay the down payment? Or do they use any collateral to get a loan? We lack some information. If yes, what could they do to make a compensation on credit or correct credit history?

We recommend legislators to have a flexible payment method for this situation. If people who have credit issues before, but who currently have enough money to make a big portion of property value and lend less loan, lenders could consider to set up a high amount of down payment standard for the people having credit issues and shorten the loan term to a period less than 30 years.

B. Comparison of approval and denied by credit by race-level

The following pivot table summarize the comparison in different race groups:

Figure 18. Summary table of comparison between approval and denied for credit by age



	approved income	credit income	approved loan	credit loan	approved property	credit property	approved loan_to_income	credit loan_to_income	approved property_to_income	credit property_to_income
derived_race										
2 or more minority races	102.092075	60.150000	311.339623	121.818182	371.026415	196.500000	3.049596	2.025240	3.634233	3.266833
American Indian or Alaska Native	99.119323	97.148148	248.386454	166.666667	306.525896	283.740741	2.505934	1.715593	3.092494	2.920701
Asian	133.309701	145.144406	422.522216	329.440559	607.650854	551.027972	3.169478	2.269743	4.558189	3.796412
Black or African American	101.069516	87.834741	322.305730	244.040307	366.319778	321.404990	3.188951	2.778403	3.624434	3.659201
Free Form Text Only	119.705882	139.250000	362.058824	155.000000	487.941176	525.000000	3.024570	1.113106	4.076167	3.770197
Joint	171.897254	124.038095	431.131589	222.904762	570.878548	298.790476	2.508077	1.797067	3.321045	2.408861
Native Hawaiian or Other Pacific Islander	100.395815	79.269231	309.911894	171.538462	397.061674	365.076923	3.086901	2.163998	3.954962	4.605531
Race Not Available	121.386910	94.726447	358.575510	228.039474	420.047608	294.560526	2.953988	2.407347	3.460403	3.109591
White	124.532937	265.779494	305.200908	171.877341	399.410956	236.801732	2.450765	0.646692	3.207272	0.890971

What surprize us are:

1.**Asian and White who have credit problems are those having higher average income.**
2.But the loan amount and property value are relatively lower than compared groups. Correspondingly, **they all got a low loan to income ratio and property to income ratio, especially in white group.**

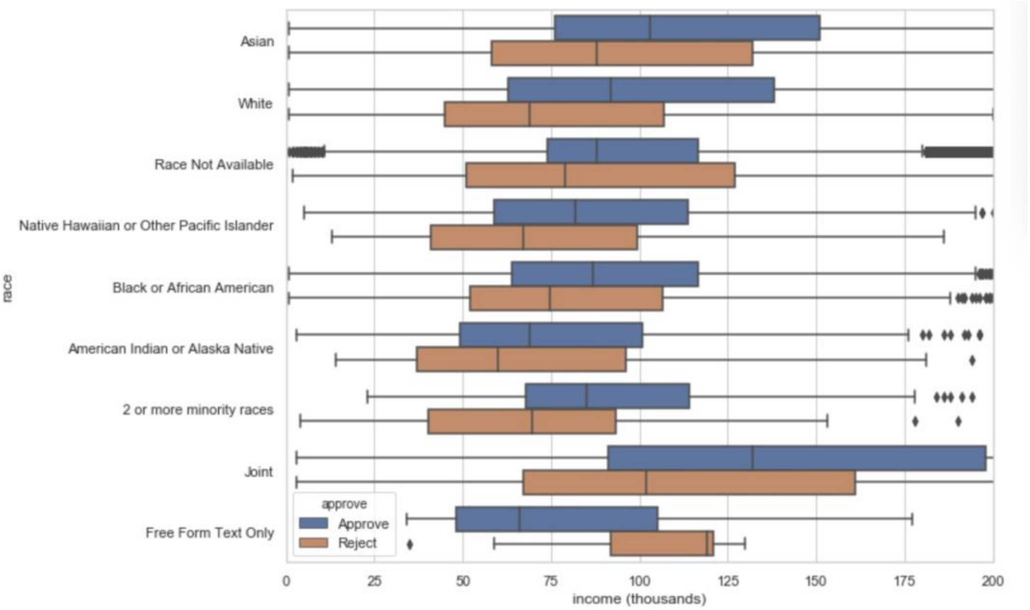
From previous exploration, we find Asian have a high level income, same as White. Unfortunately, we lack information to make a judgment on why this happens. **One reason we predict is that the income of this group of people are just getting to such a high level, in other words, their financial situation is just turning better.** They didn't have a good financial situation before so they have credit issues from the past. But their income has improved for now and they decide to purchase property. Because they don't have too much money to make down payment so that they have to purchase a relatively low value property compared to their own race people.

What we recommend here is to suggest legislators make any policy for them to prove their ability to repay the loan. Similar to the suggestions for people who have money to make down payment but have credit issues, they can be required to make a big portion of both down payment and installment in a short loan term. Let's assume, maybe 15 years, with a relatively high interest rate.

2.4 Other comparison analysis

2.4.1 Hypothesis test of loan cost in different race groups

Figure 19.Relationship between races and application status



Above figure shows the relationship between races and application status. We can observe that ‘‘White’, ‘ Native Hawaiian or Other Pacific Islander’ and ‘American Indian or Alaska Native’ have lower median income in the reject zone. On the other hand, ‘Asian’ and ‘Joint’ hold higher median income. This situation indicates the immigration might have a higher rate to be rejected based on the same income level when compared to the American. But this conclusion may be due to the bias because we have only a few data in certain races.

Additionally, we generated bootstrap replicates of the mean of the **total loan cost** of **White people and Nativie Hawaiian or other Pacific Islander** and conducted a t-test to see if there is a significant difference between them. Our hypotheses are:

Null Hypothesis: The mean total loan costs of White and Native have identical probability distribution.

Alternative Hypothesis: there is a significant difference between the mean total loan cost of White people and Nativie.

Since the p-value is 0, We reject the mean total loan costs are significantly different for White people and Native people. We are guessing there is a discrimination on loan costs based on race.

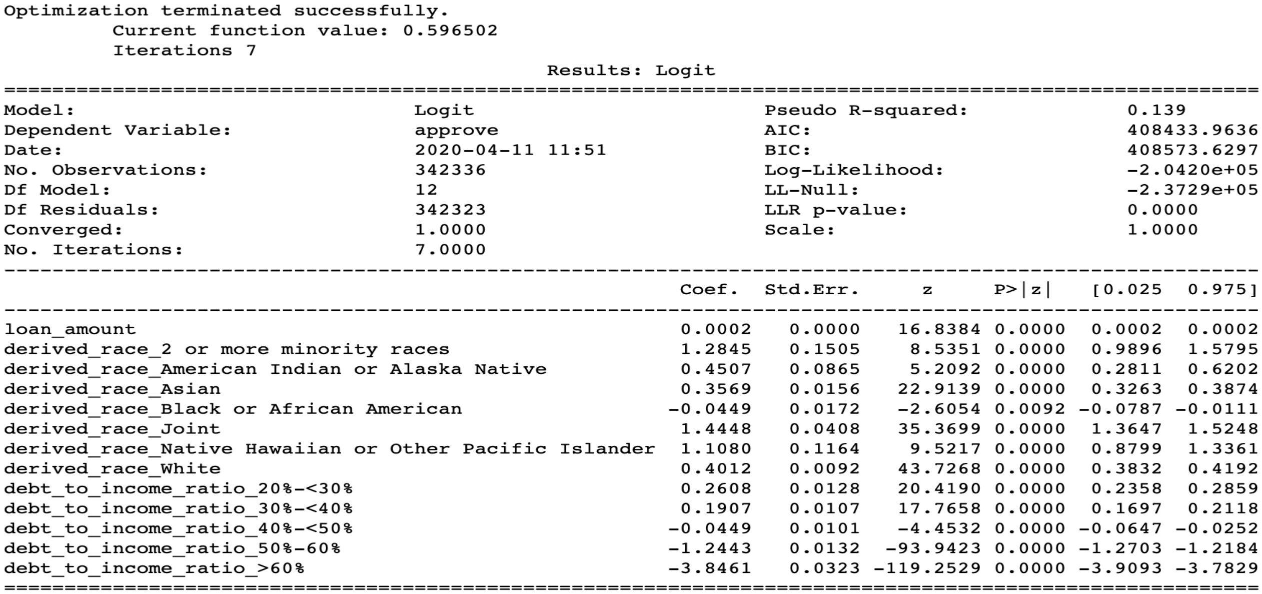
2.4.2 Logistic regression of factors affecting approval rate

From the above analysis, we have a basic idea that debt to income ratio, loan amount and race might have influence on the approval of loan. Hence, we conduct a logistic regression to test our conclusion.

Logistic regression:

Binary Response: Approve or Rejected?
Predictors: Loan amount, Debt to Income Ratio(dummy variable), Race(dummy variable)

Figure 20. Output of logistic regression



```
Optimization terminated successfully.
         Current function value: 0.596502
         Iterations 7
                                             Results: Logit
==========================================================================================================
Model:                            Logit                        Pseudo R-squared:              0.139
Dependent Variable:               approve                      AIC:                           408433.9636
Date:                             2020-04-11 11:51             BIC:                           408573.6297
No. Observations:                 342336                       Log-Likelihood:                -2.0420e+05
Df Model:                         12                           LL-Null:                       -2.3729e+05
Df Residuals:                     342323                       LLR p-value:                   0.0000
Converged:                        1.0000                       Scale:                         1.0000
No. Iterations:                   7.0000
----------------------------------------------------------------------------------------------------------
                                                        Coef.  Std.Err.      z     P>|z|   [0.025  0.975]
----------------------------------------------------------------------------------------------------------
loan_amount                                             0.0002   0.0000   16.8384 0.0000  0.0002  0.0002
derived_race_2 or more minority races                   1.2845   0.1505    8.5351 0.0000  0.9896  1.5795
derived_race_American Indian or Alaska Native           0.4507   0.0865    5.2092 0.0000  0.2811  0.6202
derived_race_Asian                                      0.3569   0.0156   22.9139 0.0000  0.3263  0.3874
derived_race_Black or African American                 -0.0449   0.0172   -2.6054 0.0092 -0.0787 -0.0111
derived_race_Joint                                      1.4448   0.0408   35.3699 0.0000  1.3647  1.5248
derived_race_Native Hawaiian or Other Pacific Islander  1.1080   0.1164    9.5217 0.0000  0.8799  1.3361
derived_race_White                                      0.4012   0.0092   43.7268 0.0000  0.3832  0.4192
debt_to_income_ratio_20%-<30%                           0.2608   0.0128   20.4190 0.0000  0.2358  0.2859
debt_to_income_ratio_30%-<40%                           0.1907   0.0107   17.7658 0.0000  0.1697  0.2118
debt_to_income_ratio_40%-<50%                          -0.0449   0.0101   -4.4532 0.0000 -0.0647 -0.0252
debt_to_income_ratio_50%-60%                           -1.2443   0.0132  -93.9423 0.0000 -1.2703 -1.2184
debt_to_income_ratio_>60%                              -3.8461   0.0323 -119.2529 0.0000 -3.9093 -3.7829
==========================================================================================================
```

For logistic regression, **we can confirm that loan amount, debt to income ratio, and race are significant factors associated with application status.** The p-values are all significant(except the race of Black or African American) and the confidence intervals are quite small. Comparing the coefficients in above table, debt to income rate and race cause more effects toward approval or not. For example, debt to income ratio greater than 60% will lead to a huge negative impact, which is also validated by the denial reason analysis above. On the other hand**, the white has some advantage compared to other races especially in American Indian and minority races**. **Hence, racial equality could be one direction for the government when designing legislation.**

Part 3 Summary and Recommendation

3.1 Summary

In order to give thorough recommendations for legislators, we perform a step by step analysis on the dataset. At first, we filtered the target applicants for the investigation. Then, from 99 variables, we select 18 of them to do the whole analysis which are relevant to loan information of applicants.

To start, we use k-means for clustering applicants by factors of loan amount, property value, income and approval status. We got three groups and find the features for each group:

- Group 1 - medium income, high approval rate, high debt ratio
- Group 2 - high income, high approval rate, lowest debt ratio
- Group 3 - low income, low approval rate ,high debt ratio

Then, we investigate loan behavior by groups according to age and race. We find that middle aged people, having relatively higher income are the majority house buyers as the box plot shows, especially for the rich. Interestingly, the elder and rich are more likely to buy high value property than others. They own a well financed situation at their age. In contrast, for elder and poor people, they can't afford to buy high value property. What's more, the Asian buy the highest value of property and their income is also relatively high in all groups. But for some people in group 3, joint race applicants are more conservative.

Next, we put lots of effort into analyzing denial reasons. Then comes a clear conclusion:

- The most common rejection reason is the debt-to-income ratio for all type applicants, followed by credit history.
- An extra finding is that low income applicants don't have enough high value assets as collateral, and they are not knowledgeable with the process of credit application.

Here, we give two recommendations for the group 3 to get an advantage in loan application. And from these conclusions, we aim to define what factors could lead applicants to have a high debt to income ratio. Then, we specify these buyers with low probability to be approved in an age-level and race-level, respectively.

We perform linear regression analysis between loan amount and property value and find a positive relationship. But we also see outliers which indicate some applicants did not behave as what we predict.

Except for property value as a reason for high debt ratio, we also assume income is also a factor. But surprisingly, income seems not a crucial factor affecting the result of approval, due to the income distribution in both approval and rejection groups being similar.

As we look at the rejection group further, we find another reason leading to a high debt ratio, which is that these people do not have a clear picture about how much they should spend on property which is proper for them financially justified. We give our suggestions to help people avoid having a high debt ratio when purchasing property.

Then, we examined approval rates in different age and race groups to give a portrait of people whose application easily gets denied for both high debt ratio and credit issue and come to four critical findings:

●People over 74 can be grouped into 2 categories: the rich with lots of deposit to purchase a house and the poor with less money but can buy a low property value house with a small amount of loan. This suggests most of this group of people is rational in purchasing property and they should get a justified evaluation on loan mortgage.

●For both of 2 or more minority races and Islander, they also have a very high property to income ratio and loan to income ratio even though their average property value is not high. This implies that their income level is too low for the current property market.

●People having credit issues and whose ages are between 55 to 64 and over 74 may have a big amount of cash on hand and borrow less due to their loan to income ratio being low and property to income ratio being high.

●Asian and White who have credit problems are those having higher average income.But the loan amount and property value are relatively lower than compared groups. Correspondingly, they all got a low loan to income ratio and property to income ratio, especially in white group.

We also find there are some differences among race groups on loan costs and approval status. Lastly, we perform logistical regression to confirm the results affecting approval rate based on the conclusions. **we can confirm that loan amount, debt to income ratio, and race are significant factors associated with application status.**

We have given thorough recommendations for the above groups of people in helping them get a justified loan. The following part will summarize all the recommendations from all perspectives.

3.2 Recommendation

Here are the list of all the recommendation we have discussed based on each finding in context:

Recommendation-1: we suggest legislators to build up a program for educating low income first time home buyers on credit evaluation, preparation and application. The way of educating could be text information, or online video link sending by email when applicants create an application account.

Recommendation-2: due to low income people have relatively less assets to be collateral. Legislators could find ways to figure out the way to be collateral other than assets. For example, legislation could require low income applicants to use financial products of deposits to be collateral, or to make a reverse mortgage to give away the right of property after an amount of years.

Recommendation-3: we recommend that legislation could require all the applicants to complete a pre approval before originating an application. Currently, pre approval is not a must have step in loan mortgage application. But, it can help first time home buyers understand the current situation they stand to make a formal decision. This way could also help them save time and money on unnecessary failures on loan application due to the loan costs.

Recommendation-4: we also believe that first time home buyers have demand on houses for families. Some low income applicants should get a proper guidance on screening property. Alternatively, the group of people who might have a high risk to be rejected on loan application should get some extra program on loan issues, like interest rate policy and payment method policy.

Recommendation-5: we recommend legislation to help people aged over 74 to get a justified consideration on loan application for property in a more specific way on interest rate, collectoral, reverse mortgage, and any other payment method. As what we discovered from our dataset, the payment ways for all applicants are the same. The results show that they didn’t use negative mortgage, interest rate payment or balloon payment. And the typical loan term for all loan applications is the same as 30 years around.

Legislator could make a more flexible loan method for this group to change some policies discussed above. One example is to use a reverse mortgage which can reduce the risk of being unable to repay the loan.

Recommendation-6: we recommend legislators to help minor races find the property based on their income level and financial situation. One blueprint is to build a community with low construction and land cost only for these minor race people to live in. The trade off might be a long travelling distance to downtown or city. Thus, this should be a social problem for the government to find a way to provide a basic place for minorities to set their homeland.

Recommendation-7: from the other perspective, legislators should also encourage relevant government departments to improve the overall education level of minorities for the next generations to increase the average level of income. Thus, it could solve the problem at the beginning.

Recommendation-8: we recommend legislators to have a flexible payment method for people who have credit issues before, but who currently have enough money to make a big portion of property value and lend less loan. Lenders could consider setting up a high amount of down payment standard for the people having credit issues and shorten the loan term to a period less than 30 years.

Recommendation-9: we suggest legislators make any policy for the Asian and the White, who have credit issues but have high income to prove their ability to repay the loan. Similar to the suggestions for people who have money to make down payment but have credit issues, they can be required to make a big portion of both down payment and installment in a short loan term. Let’s assume, maybe 15 years, with a relatively high interest rate. Or encourage them to make a reverse mortgage in case they can’t make the promise to repay the loan.

Part 4 Appendix

Figure 1. Numerical summary of loan information

Mean Income(in thousands)	$124.57
Mean Property Value(in thousands)	$416.55
Mean Loan Amount(in thousands)	$324.46
Mean Total Loan Costs(in dollars)	$5532
Mean Loan Term	347 months
Mean Interest Rate	4.59%

Figure 2. Percentage summary table of loan type

Loan Type	Percentage
Conventional (not insured or guaranteed by FHA, VA, RHS, or FSA)	73.1%
Federal Housing Administration insured (FHA)	20.2%
Veterans Affairs guaranteed (VA)	5.4%
USDA Rural Housing Service or Farm Service Agency guaranteed (RHS or FSA)	1.3%

Figure3. Bar charts of geographical information of applicants

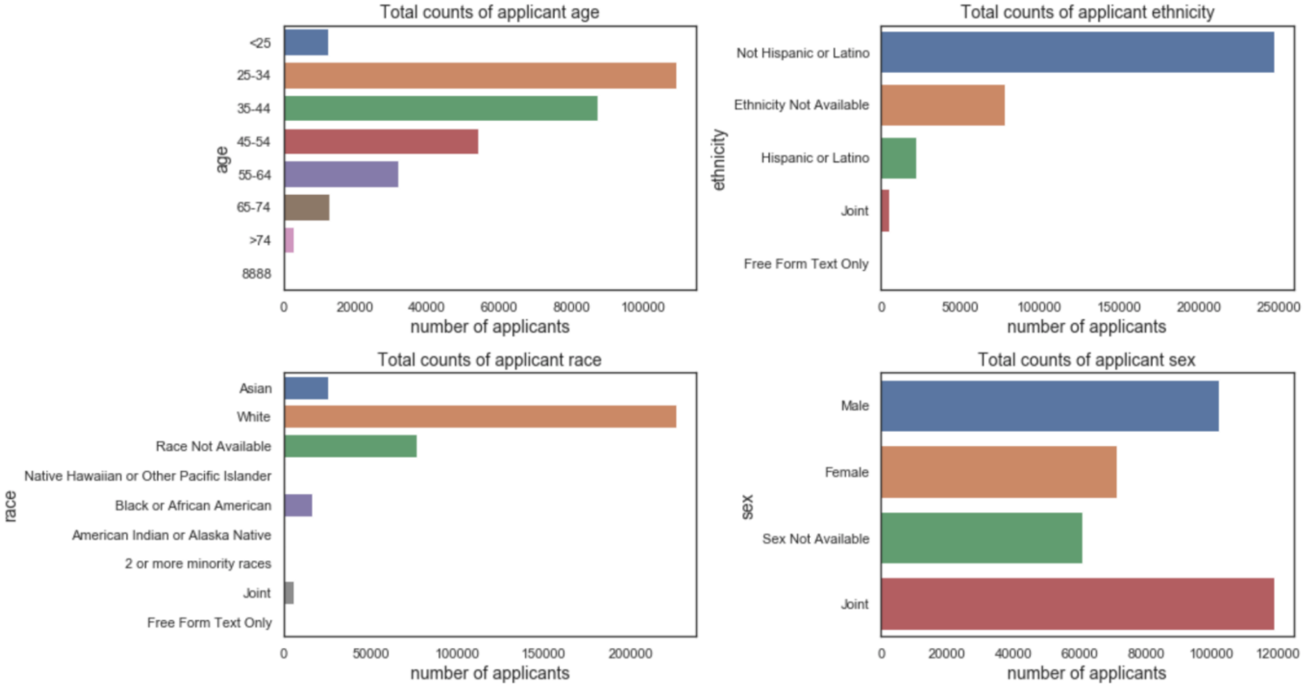


Figure 4. Number of clusters by elbow method

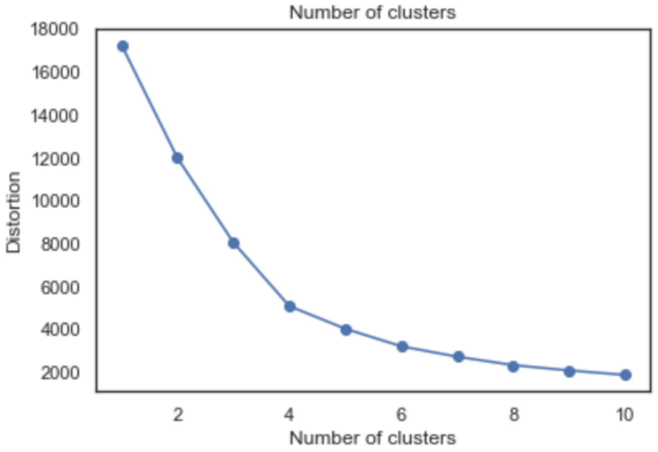
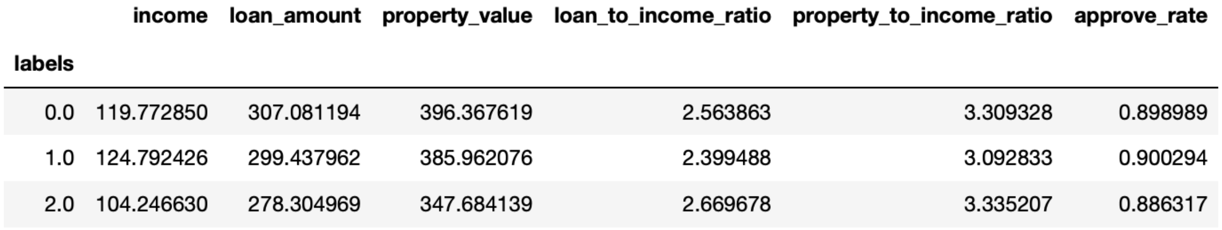


Figure 5. Classification of applicants by k-means method



labels	income	loan_amount	property_value	loan_to_income_ratio	property_to_income_ratio	approve_rate
0.0	119.772850	307.081194	396.367619	2.563863	3.309328	0.898989
1.0	124.792426	299.437962	385.962076	2.399488	3.092833	0.900294
2.0	104.246630	278.304969	347.684139	2.669678	3.335207	0.886317

Figure 6. Boxplots of income distribution by age

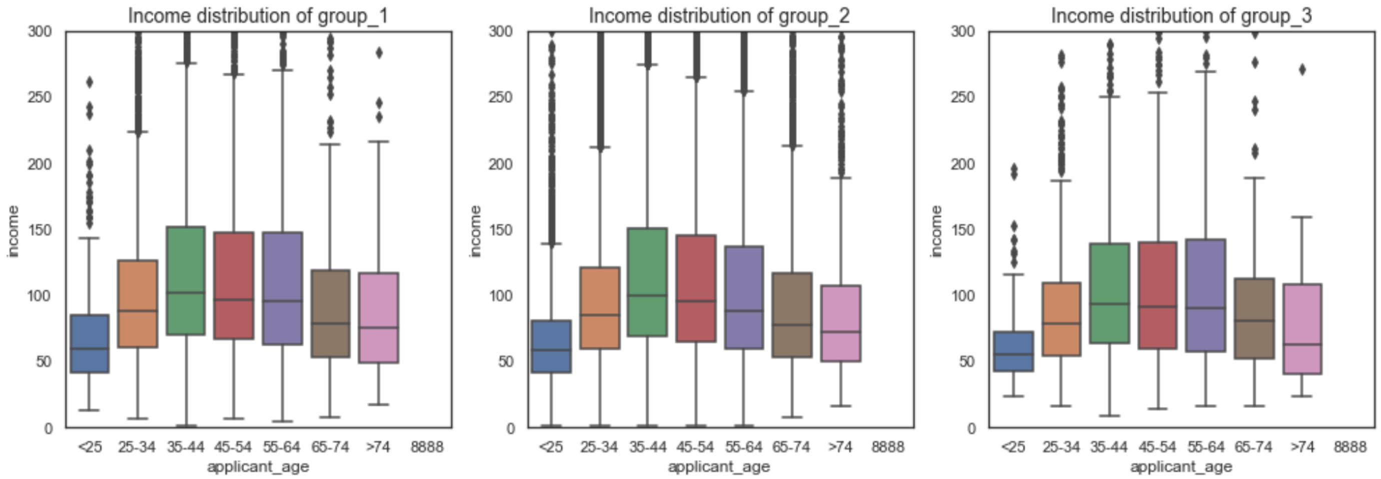


Figure 7. Boxplots of property value distribution by age

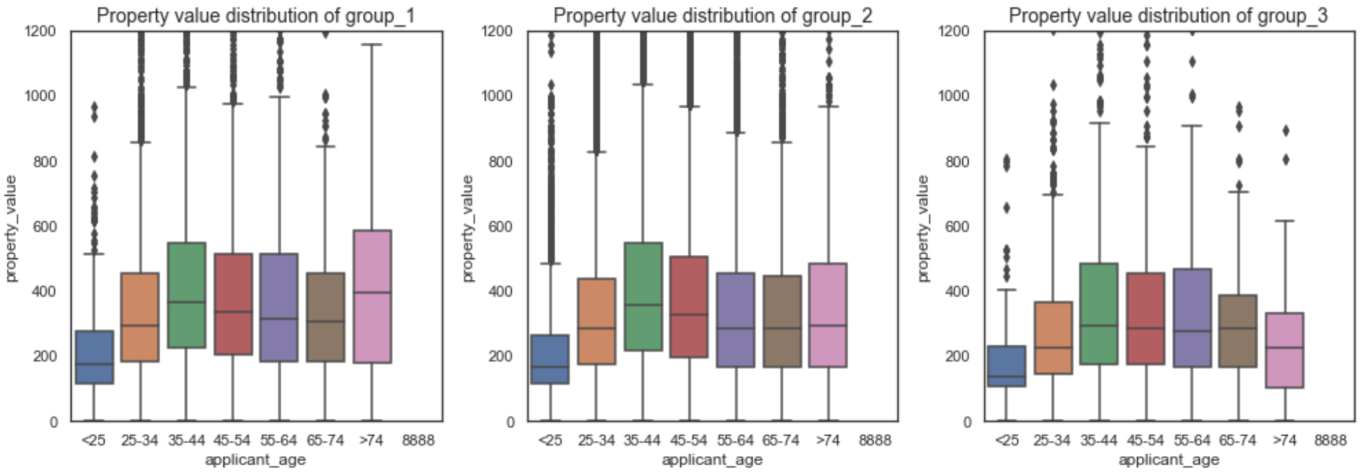


Figure 8. Box plots of income distribution by race

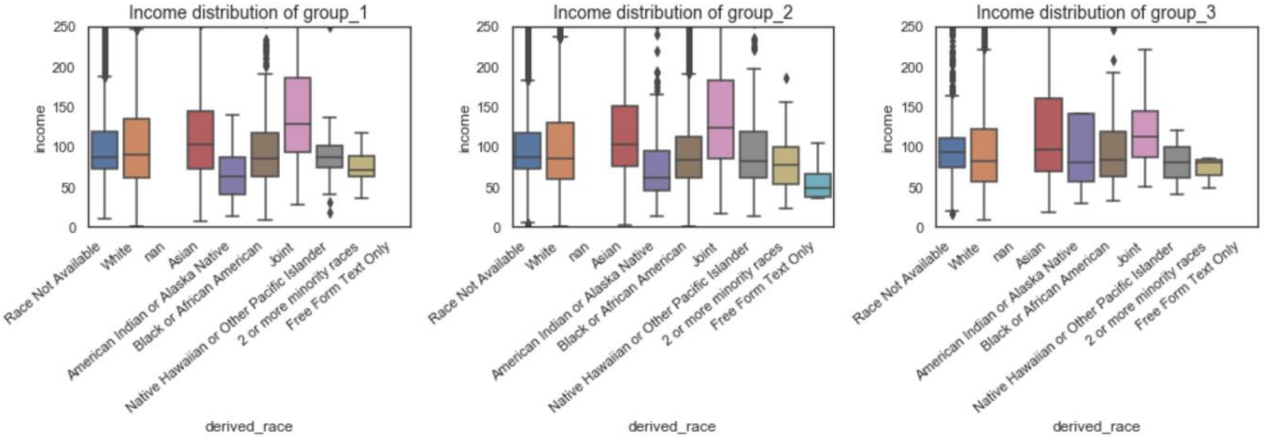


Figure 9. Box plots of property value distribution by race

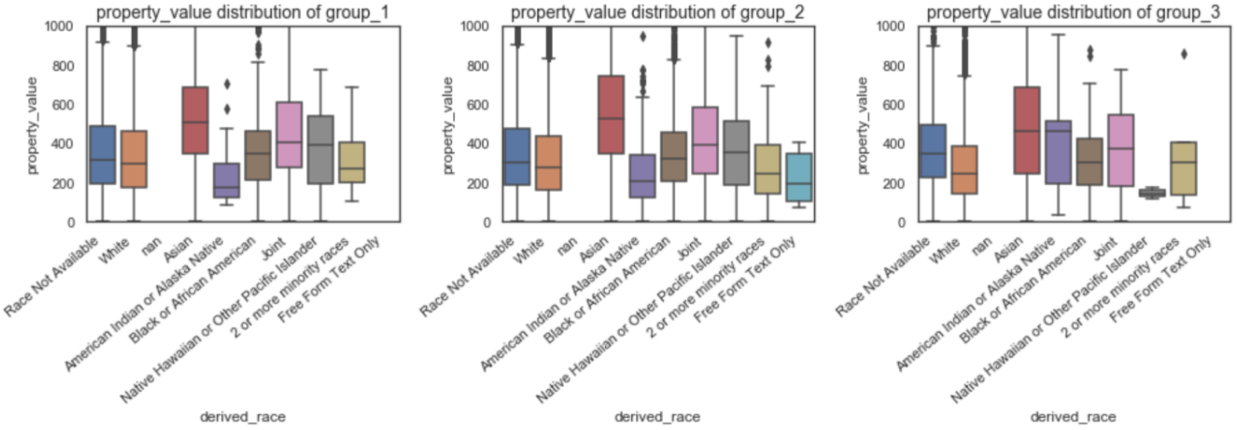


Figure 10. Bar charts about denial reason in each group

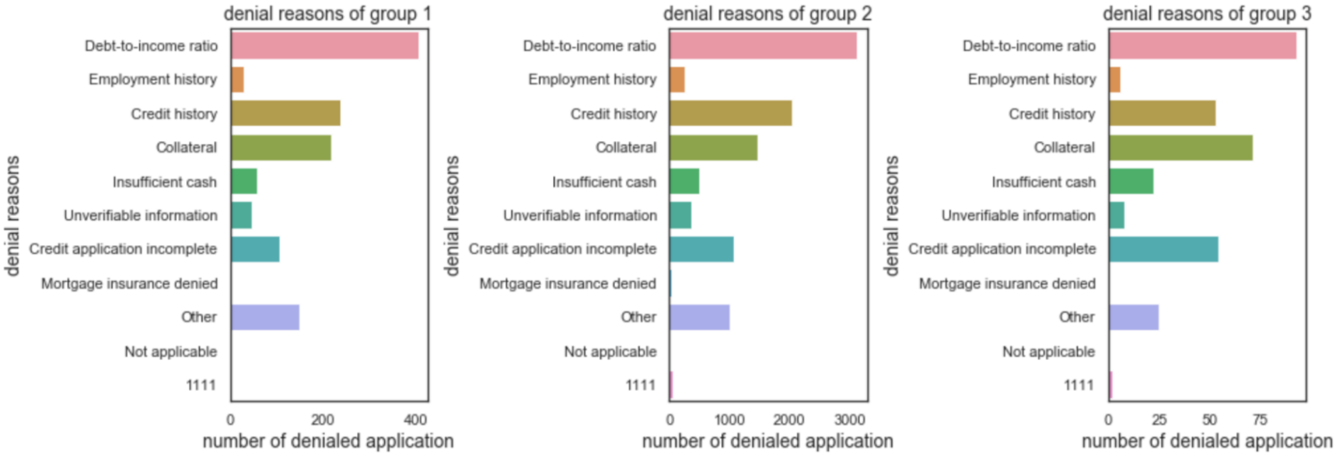


Figure 11. Scatter plot of loan and property

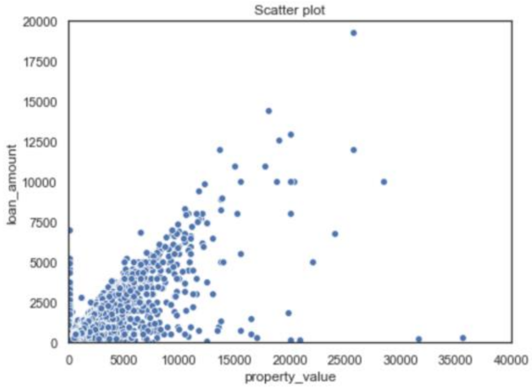


Figure 12. Linear regression of loan and property

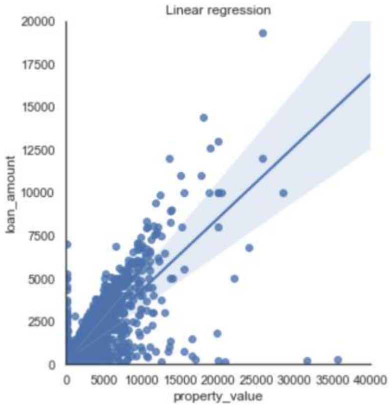


Figure 13 . Comparison of income against debt to income ratio between approval and denial

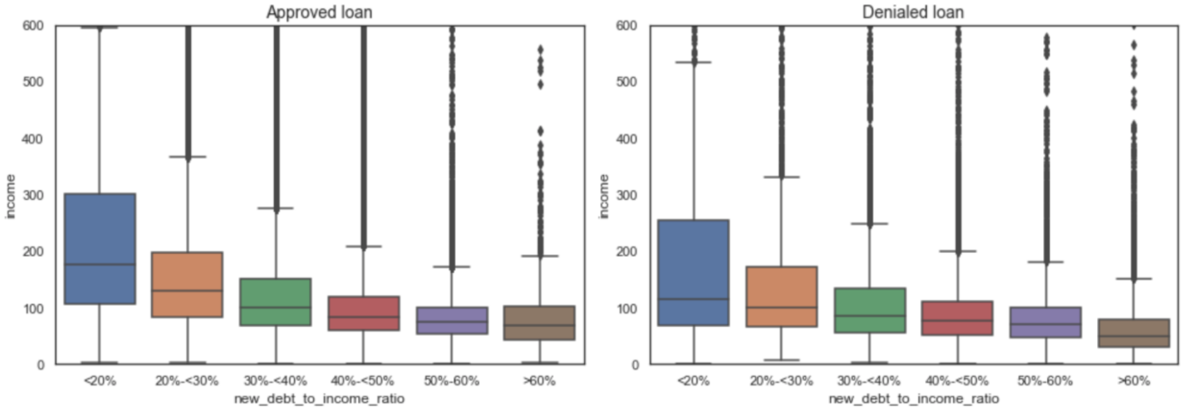


Figure 14 . Comparison of property value against debt to income ratio between approval and denial

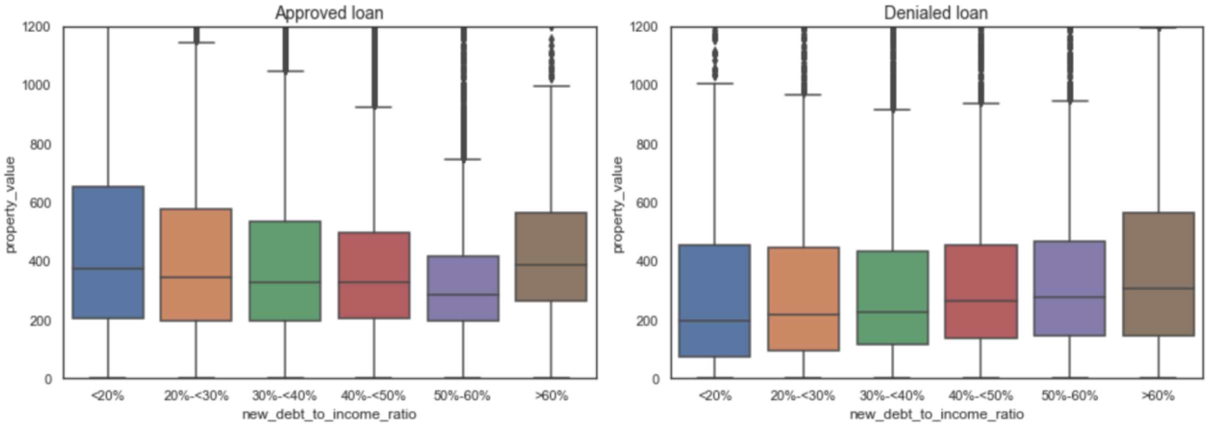
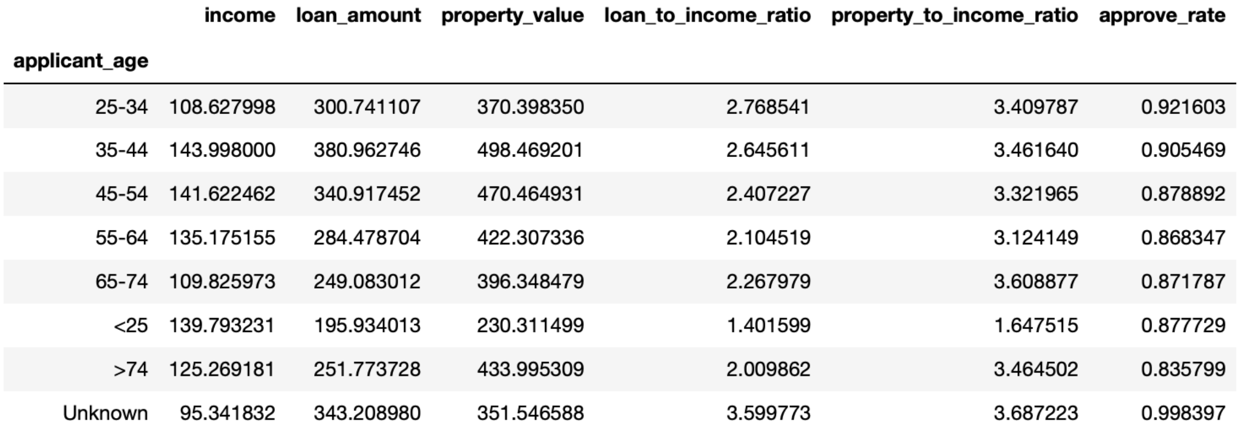
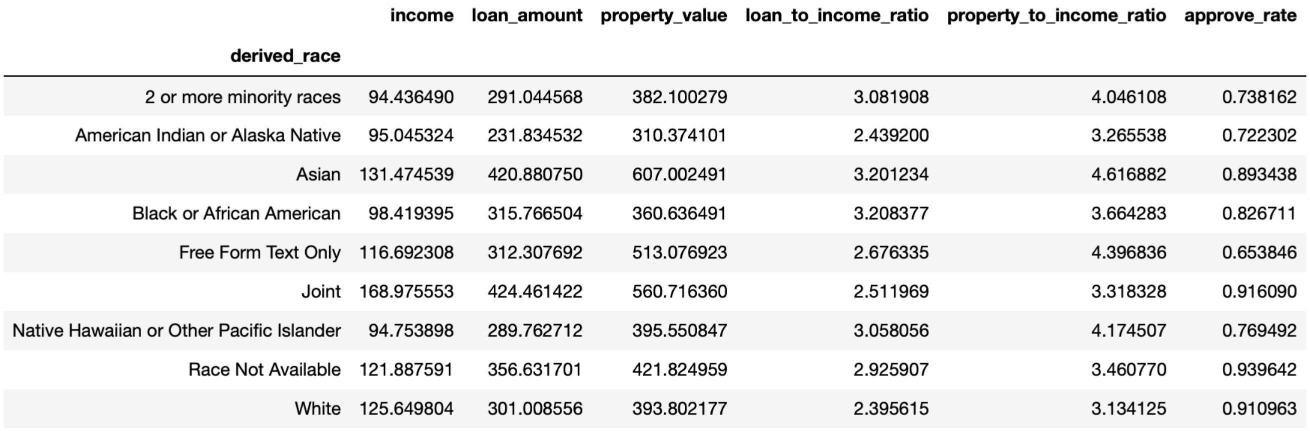


Figure 15. Summary table of mortgage ratio with approve rate by age



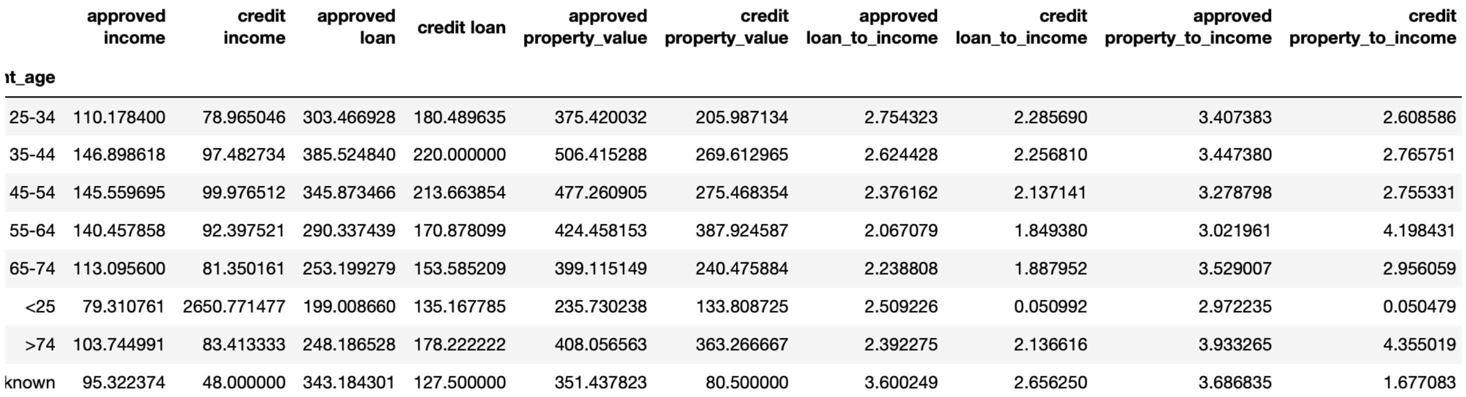
	income	loan_amount	property_value	loan_to_income_ratio	property_to_income_ratio	approve_rate
applicant_age						
25-34	108.627998	300.741107	370.398350	2.768541	3.409787	0.921603
35-44	143.998000	380.962746	498.469201	2.645611	3.461640	0.905469
45-54	141.622462	340.917452	470.464931	2.407227	3.321965	0.878892
55-64	135.175155	284.478704	422.307336	2.104519	3.124149	0.868347
65-74	109.825973	249.083012	396.348479	2.267979	3.608877	0.871787
<25	139.793231	195.934013	230.311499	1.401599	1.647515	0.877729
>74	125.269181	251.773728	433.995309	2.009862	3.464502	0.835799
Unknown	95.341832	343.208980	351.546588	3.599773	3.687223	0.998397

Figure 16. Summary table of mortgage ratio with approve rate by race



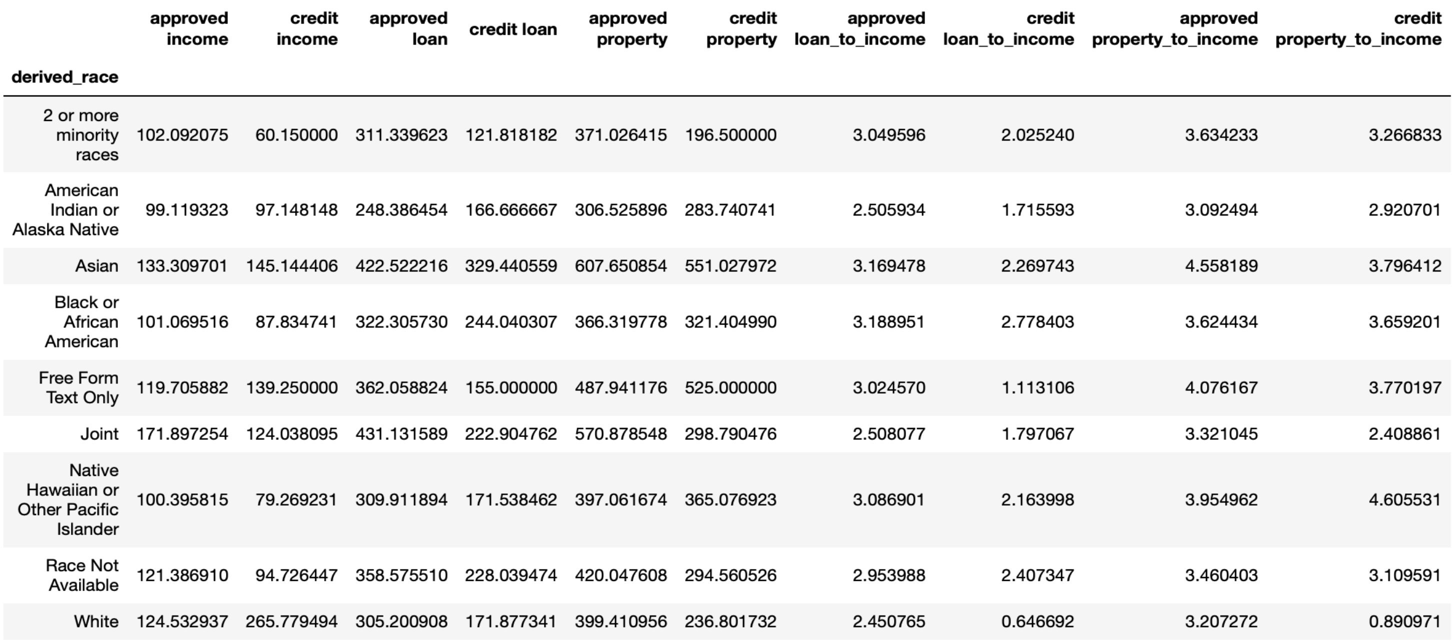
	income	loan_amount	property_value	loan_to_income_ratio	property_to_income_ratio	approve_rate
derived_race						
2 or more minority races	94.436490	291.044568	382.100279	3.081908	4.046108	0.738162
American Indian or Alaska Native	95.045324	231.834532	310.374101	2.439200	3.265538	0.722302
Asian	131.474539	420.880750	607.002491	3.201234	4.616882	0.893438
Black or African American	98.419395	315.766504	360.636491	3.208377	3.664283	0.826711
Free Form Text Only	116.692308	312.307692	513.076923	2.676335	4.396836	0.653846
Joint	168.975553	424.461422	560.716360	2.511969	3.318328	0.916090
Native Hawaiian or Other Pacific Islander	94.753898	289.762712	395.550847	3.058056	4.174507	0.769492
Race Not Available	121.887591	356.631701	421.824959	2.925907	3.460770	0.939642
White	125.649804	301.008556	393.802177	2.395615	3.134125	0.910963

Figure 17. Summary table of comparison between approval and denied for credit by age



	approved income	credit income	approved loan	credit loan	approved property_value	credit property_value	approved loan_to_income	credit loan_to_income	approved property_to_income	credit property_to_income
nt_age										
25-34	110.178400	78.965046	303.466928	180.489635	375.420032	205.987134	2.754323	2.285690	3.407383	2.608586
35-44	146.898618	97.482734	385.524840	220.000000	506.415288	269.612965	2.624428	2.256810	3.447380	2.765751
45-54	145.559695	99.976512	345.873466	213.663854	477.260905	275.468354	2.376162	2.137141	3.278798	2.755331
55-64	140.457858	92.397521	290.337439	170.878099	424.458153	387.924587	2.067079	1.849380	3.021961	4.198431
65-74	113.095600	81.350161	253.199279	153.585209	399.115149	240.475884	2.238808	1.887952	3.529007	2.956059
<25	79.310761	2650.771477	199.008660	135.167785	235.730238	133.808725	2.509226	0.050992	2.972235	0.050479
>74	103.744991	83.413333	248.186528	178.222222	408.056563	363.266667	2.392275	2.136616	3.933265	4.355019
known	95.322374	48.000000	343.184301	127.500000	351.437823	80.500000	3.600249	2.656250	3.686835	1.677083

Figure 18. Summary table of comparison between approval and denied for credit by age



	approved income	credit income	approved loan	credit loan	approved property	credit property	approved loan_to_income	credit loan_to_income	approved property_to_income	credit property_to_income
derived_race										
2 or more minority races	102.092075	60.150000	311.339623	121.818182	371.026415	196.500000	3.049596	2.025240	3.634233	3.266833
American Indian or Alaska Native	99.119323	97.148148	248.386454	166.666667	306.525896	283.740741	2.505934	1.715593	3.092494	2.920701
Asian	133.309701	145.144406	422.522216	329.440559	607.650854	551.027972	3.169478	2.269743	4.558189	3.796412
Black or African American	101.069516	87.834741	322.305730	244.040307	366.319778	321.404990	3.188951	2.778403	3.624434	3.659201
Free Form Text Only	119.705882	139.250000	362.058824	155.000000	487.941176	525.000000	3.024570	1.113106	4.076167	3.770197
Joint	171.897254	124.038095	431.131589	222.904762	570.878548	298.790476	2.508077	1.797067	3.321045	2.408861
Native Hawaiian or Other Pacific Islander	100.395815	79.269231	309.911894	171.538462	397.061674	365.076923	3.086901	2.163998	3.954962	4.605531
Race Not Available	121.386910	94.726447	358.575510	228.039474	420.047608	294.560526	2.953988	2.407347	3.460403	3.109591
White	124.532937	265.779494	305.200908	171.877341	399.410956	236.801732	2.450765	0.646692	3.207272	0.890971

Figure 19. Relationship between races and application status

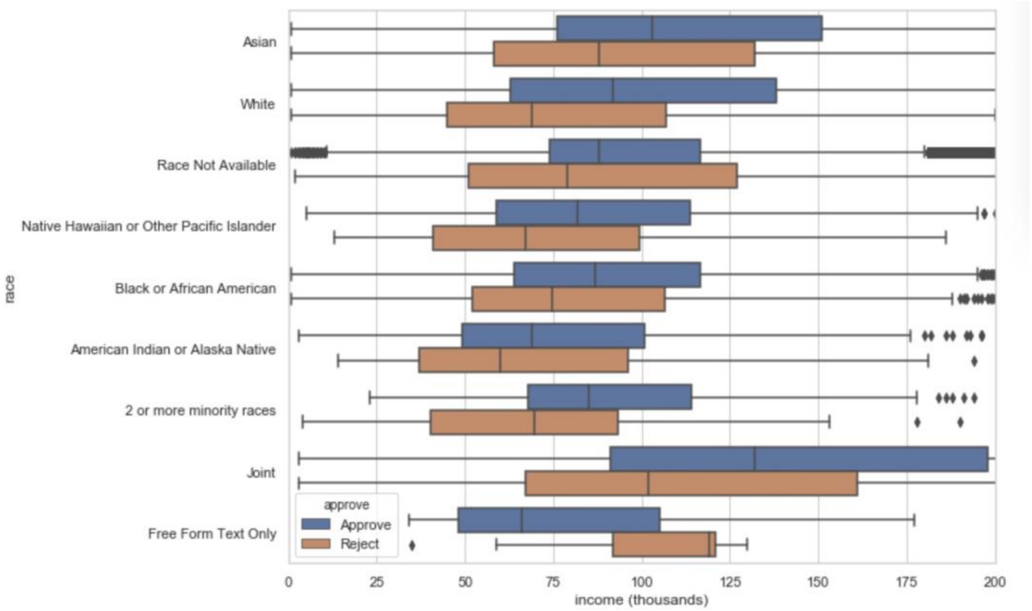
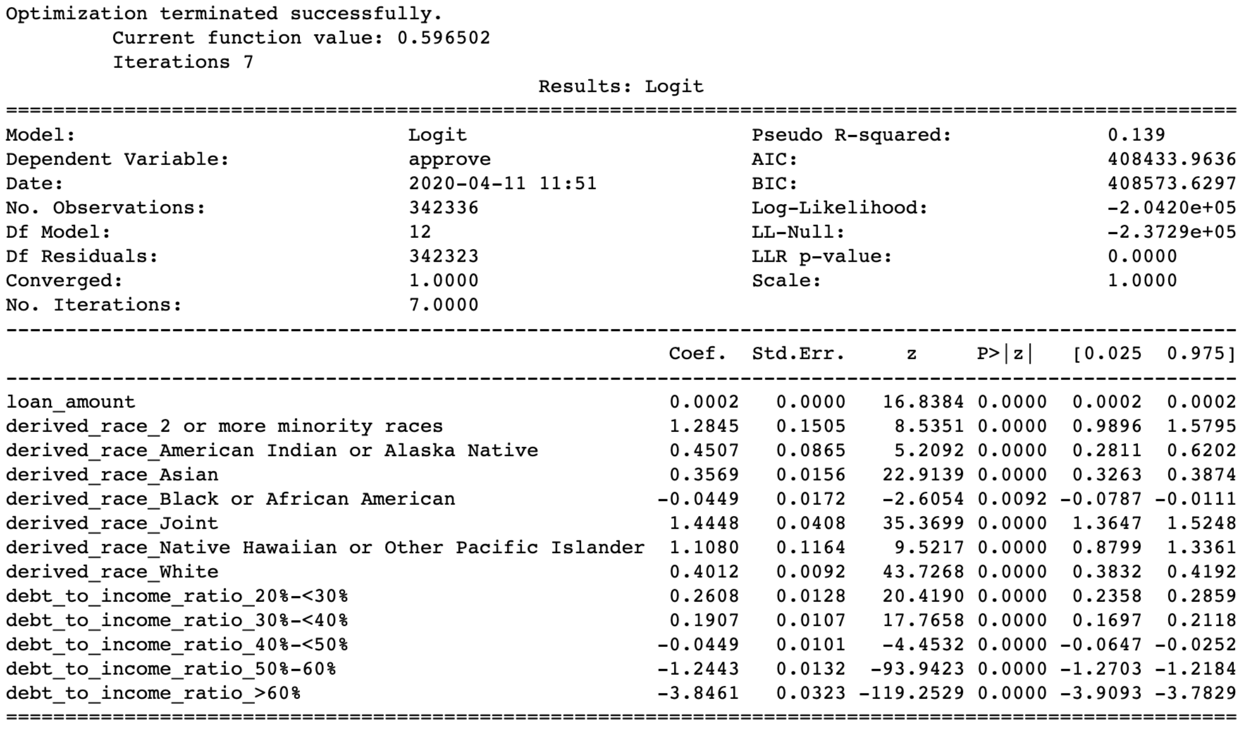


Figure 20. Output of logistic regression



```
Optimization terminated successfully.
         Current function value: 0.596502
         Iterations 7
                                         Results: Logit
=========================================================================================================
Model:                           Logit                         Pseudo R-squared:              0.139
Dependent Variable:              approve                       AIC:                           408433.9636
Date:                            2020-04-11 11:51              BIC:                           408573.6297
No. Observations:                342336                        Log-Likelihood:                -2.0420e+05
Df Model:                        12                            LL-Null:                       -2.3729e+05
Df Residuals:                    342323                        LLR p-value:                   0.0000
Converged:                       1.0000                        Scale:                         1.0000
No. Iterations:                  7.0000
---------------------------------------------------------------------------------------------------------
                                                        Coef.  Std.Err.     z     P>|z|   [0.025  0.975]
---------------------------------------------------------------------------------------------------------
loan_amount                                             0.0002   0.0000   16.8384 0.0000  0.0002  0.0002
derived_race_2 or more minority races                   1.2845   0.1505    8.5351 0.0000  0.9896  1.5795
derived_race_American Indian or Alaska Native           0.4507   0.0865    5.2092 0.0000  0.2811  0.6202
derived_race_Asian                                      0.3569   0.0156   22.9139 0.0000  0.3263  0.3874
derived_race_Black or African American                 -0.0449   0.0172   -2.6054 0.0092 -0.0787 -0.0111
derived_race_Joint                                      1.4448   0.0408   35.3699 0.0000  1.3647  1.5248
derived_race_Native Hawaiian or Other Pacific Islander  1.1080   0.1164    9.5217 0.0000  0.8799  1.3361
derived_race_White                                      0.4012   0.0092   43.7268 0.0000  0.3832  0.4192
debt_to_income_ratio_20%-<30%                           0.2608   0.0128   20.4190 0.0000  0.2358  0.2859
debt_to_income_ratio_30%-<40%                           0.1907   0.0107   17.7658 0.0000  0.1697  0.2118
debt_to_income_ratio_40%-<50%                          -0.0449   0.0101   -4.4532 0.0000 -0.0647 -0.0252
debt_to_income_ratio_50%-60%                           -1.2443   0.0132  -93.9423 0.0000 -1.2703 -1.2184
debt_to_income_ratio_>60%                              -3.8461   0.0323 -119.2529 0.0000 -3.9093 -3.7829
=========================================================================================================
```