Home Loan Racial Inequality

Kenny Tang, Sara Haptonstall and Nick Wu

Background

Recently, our team came across an article published by Reveal [1]news, that talks about the presence of discrimination in home mortgage loans today. In their analysis, they were able to determine the likelihood of mortgage denials for different minority groups using binary logistic regression. The results of their research showed large significant discrepancies across 48 different metropolitan areas.

This research was conducted over 5 years ago using 2015 and 2016 loan application data provided by Home Mortgage Disclosure Act (HMDA). As years past and America approaches a more multiculturally accepting society, our team is curious to see how discrimination affects the mortgage loan market today. At the time of our analysis, it is the year 2022 and the home mortgage data we will be analyzing will cover loans approved and rejected in 2021, In addition to this, we used a dataset provided by Home Owner Loan Corporation (HOLC) to identify areas that may be most affected by discrimination.

What is redlining?

The term redlining comes from the practice of the FHA and the HOLC of color-coding neighborhood maps in order to identify how safe it was to insure mortgages in certain neighborhoods. Redlining is the practice of refusing a loan to someone who is otherwise credit worthy due to the location in which the loan is applied for. These areas are marked red by loan institutions, hence the name redlining.

Redlining was prevalent for decades since 1930 and even after it was long outlawed when the Fair Housing Act and the Equal Credit Opportunity Act were enacted, it is suspected to have lingering effects.

HOLC graded neighborhoods by a scale of A to D in which: A- “Best” (Green), B- “Still Desirable” (Blue), C- “Defiantly Declining” (Yellow), D- “Hazardous” (Red). These classifications were primarily racially motivated, labeling neighborhoods with high proportion of minorities in the C and D category. Historically, areas with low grades suffer the most as mortgage companies prefer not to approve loans to these regions and because applicants who apply for loans in these neighborhoods were predominantly minorities, it was later deemed discriminatory

Project Goals

Our team proposes to examine credit data made available by the Home Owners’ Loan Act (HOLA), census data and historical HOLC neighborhoods ratings to explore the relationship between home mortgages and race in 2021 hoping to answer the following questions:

Are there disparities in the rate of interest charged to different racial groups?

Given similar interest rates, are some racial groups denied mortgages over others?

What are the differences that prevail in areas previously redlined?

In 1968, the Fair Housing Act was passed, which makes it unlawful to discriminate in terms or conditions in the basis of race or national origin.

And in 1974, The Equal Credit Opportunity Act (ECOA) enacted unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction, based on race, color, religion, national origin, sex, marital status, or age.

1. https://revealnews.org/article/how-we-identified-lending-disparities-in-federal-mortgage-data/
2. https://www.hud.gov/program_offices/fair_housing_equal_opp/fair_housing_act_overview

Data Sources

National Loan application records (LAR) 2021

·Home loan applications reported by lending institutions.
·Format: CSV
·Access method: Download

We will focus on a subset that represent applicants for single family homes what intend to use the property as their primary residence without business purposes, variables will be **filtered** as follow:

·**Loan Type:** Only Conventional: First Lien
·**Home type:** Only Single Family (1-4 units) both Site Built and Manufactured.
·**Application Outcome:** We will exclude #4 - Application withdrawn by applicant and #5-File closed for incompleteness
·**Loan purpose:** We will focus on home purchase.
·**Business or commercial purpose**: Exclude #1 for business or commercial purposes.
·**Occupancy Type**: Only #1-Principal residence
·**lien status:** Only first lien loans.
·**Income:** Keep income between $20,000 and $500,000.
·**Interest Rate:** Filter out interest rate of 100% (Since, this is implausible)
·**Census Tract:** Drop any records without census tract number without census tract, we are unable to identify the geographical region of the property.
·**Exclude:** reverse mortgages, open line of credit, negative amortization, balloon payments, in order to reduce the number of variables that may impact loan amount.

Data Manipulations:
Derived Race or Ethnicity: Keep only records with races/ethnicity: White, Black, Asian and Hispanic. Create a single columns variable, by merging Hispanic, which is recorded as ethnicity and races.

Approval Rate: Approved application divided by total applications.

Home Owner Loan Corporation (HOLC)

- Historical HOLC neighborhood ratings (A, B, C, D) U.S. cities neighborhoods Size: 1.86 MB
- Format: CSV
- Access method: Download

Data Manipulations:
FIPS code: obtained 5-character prefix from census tract number corresponding to county code (ref. page 3)
Area rated: Calculate percentage of rated area divide area (A, B, C, D) between total area rated.
Historical Redlining Score: Weighted census tract score (ref.page 4 for detailed calculations)

Census Data

- 2015-2022 Census data.
- Format: Pandas DataFrame.
- Access Method: API.

Data Manipulations:
FIPS code: Extract county code from census tract
Vacant percentage: Divide vacancy between total houses.
Mortgage percentage: One minus quotient of number of homes without a mortgage between total houses.
Minority percentage: One minus quotient of number of white people between total population.

Data Sources

County Code Geolocation.

- County FIPS code, latitude, longitude, county name
- Format: CSV
- Access method: Download

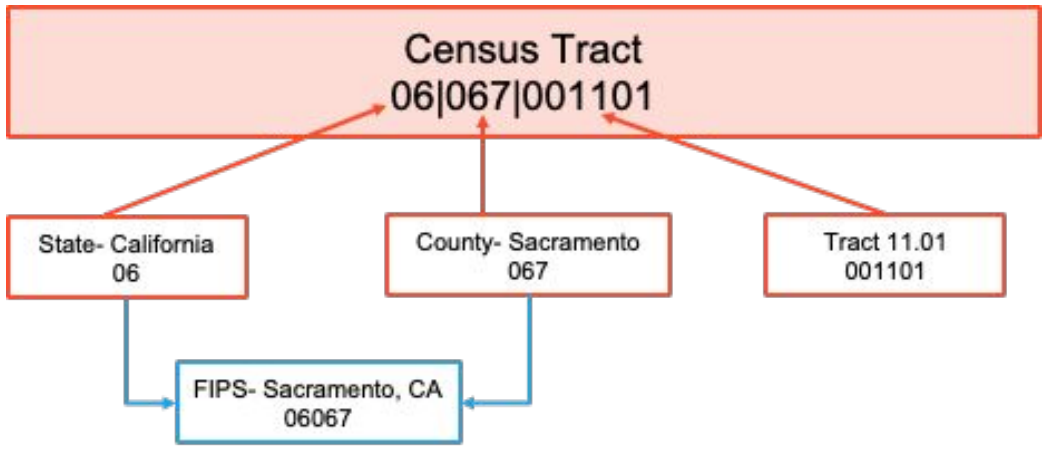
Merge FIPS code with HOLC data set for map location.

We used either the FIPS code or census tract to identify specific regions in our dataset in order to created both Dataset 1 (an individual records dataset) and Dataset 2 (an aggregated dataset).

Using redline grades and region lines from 1930s and region lines derived from the 2020 Census, we computed new Historic Redlining Scores representing a 2020 map of the United States. Our first Dataset was used to analyze the United States as a whole and the second dataset was used to examine specific counties.

Methodology

Using the census tract from our data sets, we were able to merge datasets by state, county, and tract ID, effectively giving each instance of loan a redlining score based on the area the loan were applied for. The census tract number can be decoded as follow:



Loan application records (LAR) 2021.
Join by: Census Tract

Home Owner Loan Corporation (HOLC)
Join by: Census Tract

Census Data
Join by: Census Tract

County Code Geolocation.
Join by: FIPS

Dataset 1:
- Individual loan applications (LAR).
- Historic Redlining grade area.
- Calculated field: HRS

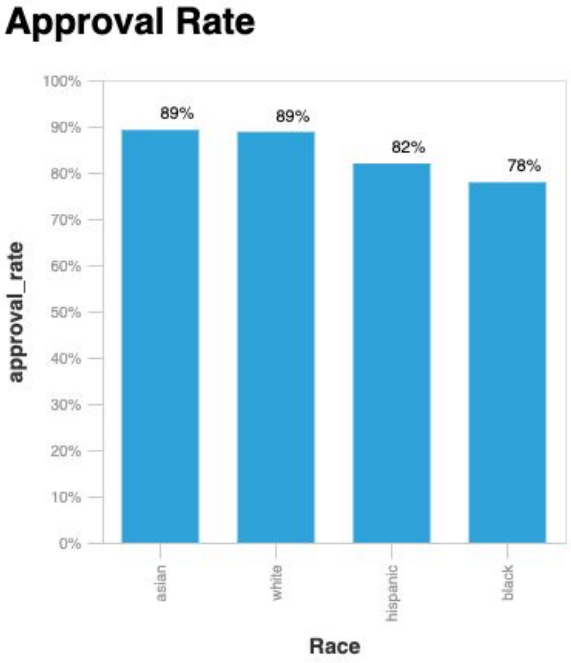
Dataset 2:
- Aggregated by census tract and county,
- loan applications data,
- HRS,
- census data (Population, number of households, minority percentage and vacant), county longitude and latitude.

Approval Rate and Historical Redlining Score (HRS)

Approval Rate by Race

Using a sample of 454,321 applications, we estimated the approval rate by race. By taking the amount of approved loans divided by the total amount of applications for each group, we were able to compute the proportion of approved loans.

Asian group has the highest, sitting at around **89% approval rate**. White group follows closely with a similar number. **Hispanic** group has a slightly **lower** approval rate of **about 82%** and the **Black** group comes last with the **lowest rate** of about **78%**, 10% lower than the White group.



Historical Redlining

Since the redlining grades were created over 80 years ago, one of the challenges we faced was that current demographic boundaries (census tracts) do not precisely align with historical redlined areas, as regions divided into counties over the years. Modern **counties in 2022, could contain multiple regions from the 1930s map** that were previously identified as grade 'A' and 'B'.

To overcome this obstacle we decided to implement a **Historic Redlining Score.** This quantifiable score was derived using the same **methodology applied in "Tracing the Legacy of Redling"** [3]

How to calculate Historical Redlining Score

Calculate total **area** A, B, C, D **proportion by census tract** listed in the HOLC data set.

Area Proportion

Calculate weighted grades.

Multiply graded area by a **weighting factor.**

- **A** or "Best" = 1
- **B** or "Desirable" = 2
- **C** or "Declining" =3
- **D** or "Hazardous" = 4

Census tract with a **score** closer to **1.00** are '**low redlining**'

Scores closer to **4.0** are **'high redlining'.**

Score

Example:

Census tract	Total Area	Rated Area	Unrated Area	HRS
1073000100	7,549,581	73.3%	26.7%	**3.53**

Rated Area	A	B	C	D
Percentage Rated	0.0%	4.3%	26.1%	42.9%
Area Proportion	(0% / 73.3%)	(4.2%/73.3%)	(26.1%/73.3%)	(42.9%/73.3%)
Weighted Grade	0 X 1	5.8 X 2	35.6 X 3	58.6 X 4
HRS	0.00	0.12	1.07	2.34

3.. https://ncrc.org/redlining-score/

Redlining

Using HOLC historical redlining information from the 1930's, we **plotted the percentage rated area by 2020 county lines** to identify current county regions where redlining occurred.

By observing the map, we can see that **redlining occurred primarily near urban areas.** There is a higher concentration in the North-East and Southern parts of the United States with most of the West and Midwest remaining ungraded. Reviewing the the demographics by county grade, we observe that **37.63%** of the current US population inhabits areas that are **graded as 'C'** .

The percentage of vacant properties for **areas graded 'C' and 'D'** are about **9% larger** than areas graded 'A'. In addition, the average proportion of **minorities living in areas graded 'D' is 6% larger** than areas graded 'A'.

By observing the average percent of minority and vacant homes for each group, we can infer a negative relationship between average minority and vacancy percentage. Areas with low grade are having difficulty filling vacant homes. Whether or not this difficulty is caused by the act of redlining, minorities are largely affected as they tend to inhabit these areas.

Next, we dive deeper into a region that has been known to have been historically redlined. To do this, we broke down a county by its census tract to compute HRS scores across the county. We then drew the map using the HRS scores computed.

The county we decided to observe redlining is Los Angeles, California. Los Angeles is one of the largest counties in the US with one of the most diverse populations.

Historical Redlining by 2020 County Lines

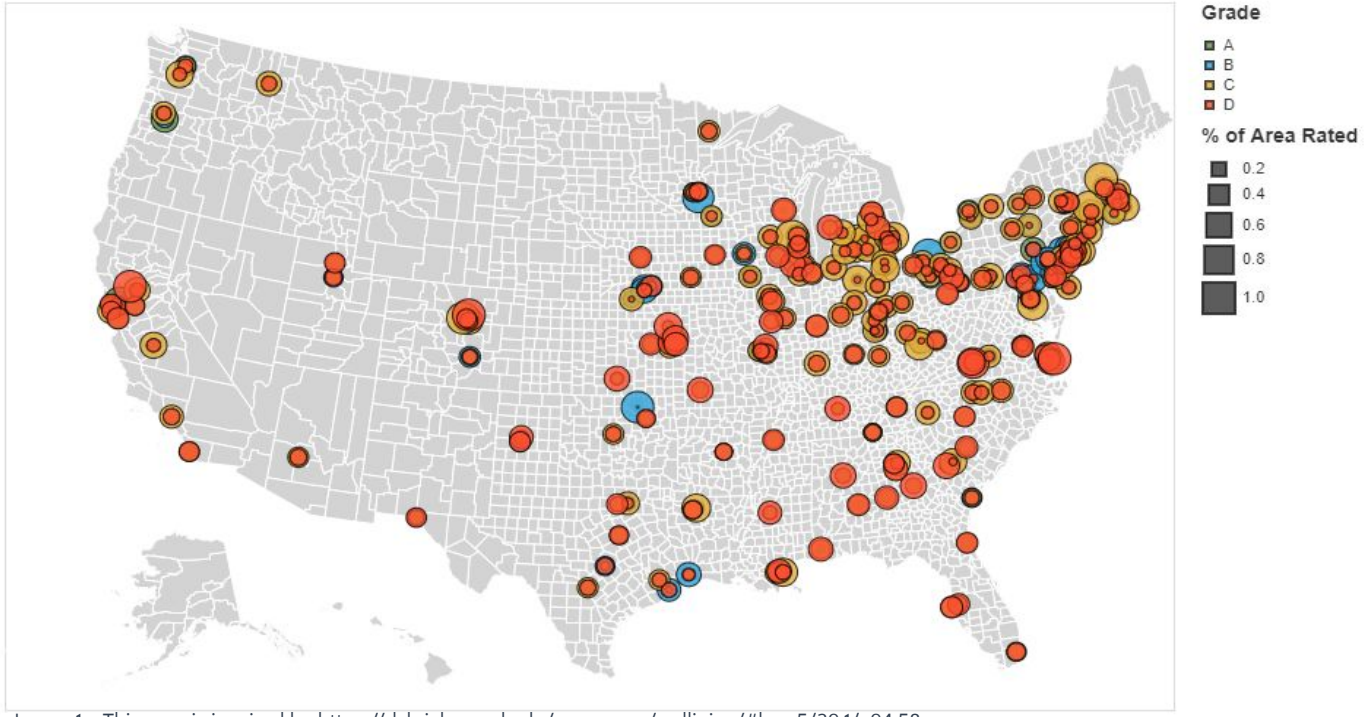


Image 1: This map is inspired by https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58

Demographics by County Grade

Grade	Population %	Population Total	Total Houses	Mean % mortgage	Mean % of Homes with Mortgages	Mean % vacant
Ungraded	54.81%	175,453,047	44,401,875	52%	16%	37%
A	0.46%	1,485,506	350,708	73%	25%	9%
B	3.85%	12,337,756	3,058,909	67%	22%	14%
C	37.63%	120,461,335	25,482,747	66%	27%	18%
D	3.24%	10,360,450	2,275,562	67%	31%	20%
Total /Mean	**100%**	**320,098,094**	**75,569,801**	**65%**	**24%**	**20%**

Los Angeles County

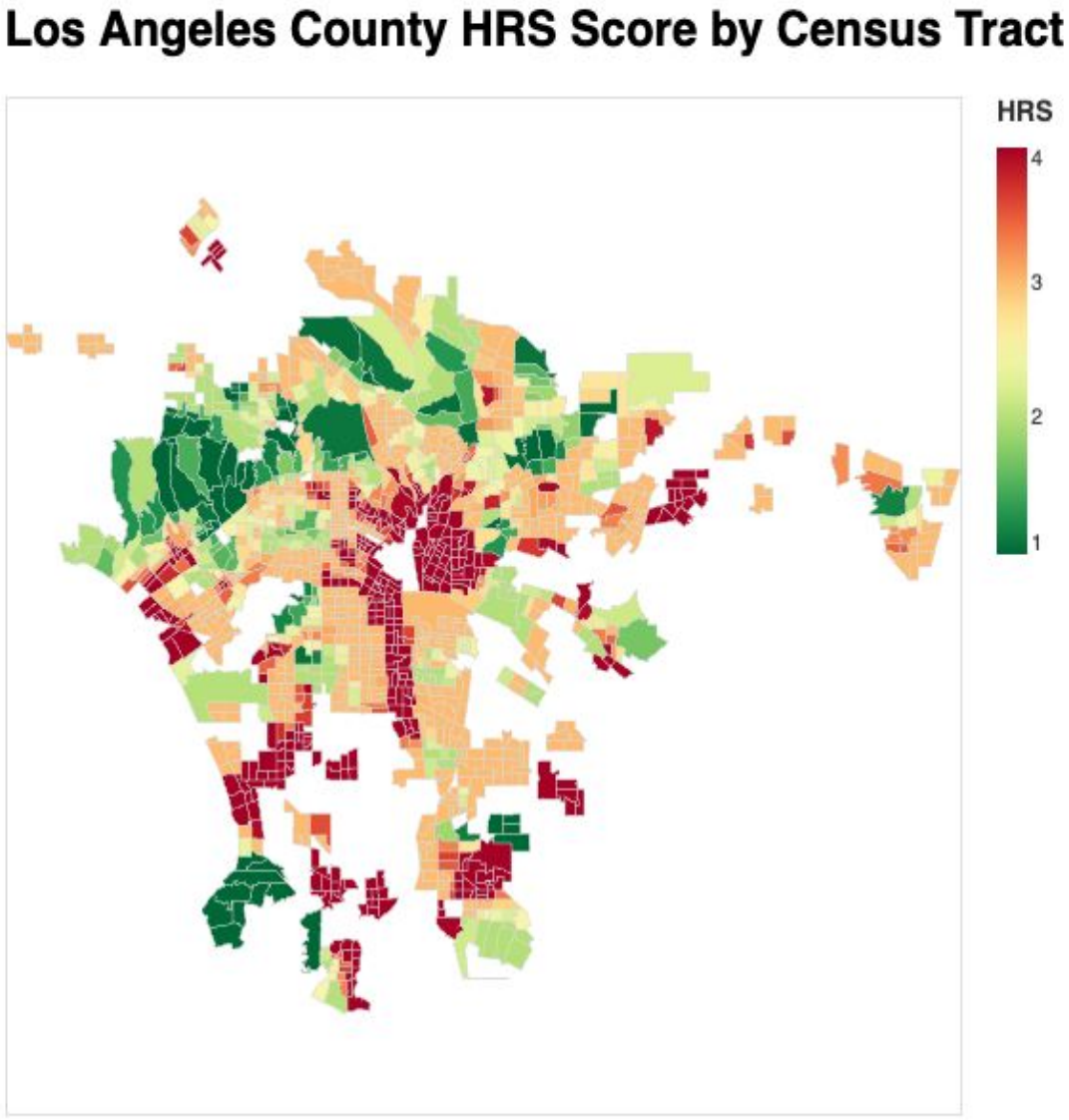
We will explore if the conditions differ significantly based on HRS by census tract in Los Angeles, California. Our calculated HRS scores range from 1.0 to 4.0, areas that were not graded were assigned a '0' value and left blank in our map. We binned our demographics by HRS into 5 bins, ungraded, [1-1.75], [1.76-2.49], [2.5-3.3] and [3.3 or more]; these correspond to the grades ungraded, A, B, C, D.

The census demographics show that **83% of the population** lives in census tracts **graded 'C' or 'D'** and the number of families exceeds the number of houses available in census tracts graded 'C' or 'D', indicating cases where **multiple families are habitating a single home**. In addition, the **mean percentage of vacant properties for grade 'C' and 'D'** is more than **doubled** that of 'B'.

We previously saw a trend of high vacancy rates in regions with low grade. For **areas graded 'D'** in **Los Angeles**, about **59%** of the 98,885 homes **were vacant,** meaning that about 58,342 homes were not in use. There are over 200,000 families living in this region, about 40,000 homes in use, and 58,000 homes not in use. We could not help but wonder if the amount of vacant homes was due to minorities having difficulty in obtaining mortgages. We decided to use a logistic regression to see if the probability of getting a mortgage loan approved in Los Angeles differ between racial groups.

Los Angeles Demographics grouped by HRS Score

		A	B	C	D
Variables	Ungraded	HRS [1 - 1.75]	HRS. [1.75-2.49]	HRS [2.5 - 3.3]	HRS [>3.3]
Population %	1%	3%	13%	55%	28%
Population Total	53,481	102,030	467,991	1,984,325	999,406
Total Houses	15,284	25,163	75,724	195,692	98,885
No. of families	14,366	24,573	105,591	419,558	207,666
Mean % of Homes with Mortgages	71%	70%	73%	75%	73%
Mean % Minority	25%	40%	40%	50%	51%
Mean % Vacant Homes	10%	16%	24%	51%	59%



Logistic Regression of Los Angeles County

Design of our Logistic Regression

To observe the differences in approval rate for the different racial groups inhabiting Los Angeles County, we ran a binary logistic regression. The dependent variable of our regression model is 'declined' a binary variable that is either 1 for denied loan or 0 for approved loan. The independent variables used for our model are log of income, log of loan amount, debt to income ratio and 4 binary variables: married applicants, asian, black, and hispanic.

Logistic Regression Results

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-11.5200	2.269	-5.078	0.000	-15.966	-7.074
asian	-0.2567	0.156	-1.648	0.099	-0.562	0.049
black	0.4107	0.266	1.543	0.123	-0.111	0.932
hispanic	0.2897	0.143	2.021	0.043	0.009	0.571
joint	-0.4290	0.131	-3.271	0.001	-0.686	-0.172
ln_income	0.5859	0.105	5.566	0.000	0.380	0.792
ln_loan_amount	-0.2800	0.187	-1.494	0.135	-0.647	0.087
debt_to_income_ratio	0.1291	0.009	14.054	0.000	0.111	0.147

Interpretation of Logistic Regression

To interpret the results, we take Euler's number 'e' to the power of the coefficient of the variable we are interested in. In our case, we are interested in the 3 minority dummy variables. By taking e to the power of the Black coefficient, we get 1.5, meaning that Black applicants were about **50% more likely** than White applicants to be denied a loan for the Los Angeles County. By performing the same computation with Hispanic and Asian coefficients, we learn that **Hispanic** applicants were about **34% more likely** than White Applicants to to be denied a loan and **Asian** applicants were about **23% less likely** than White applicants to be denied a loan.

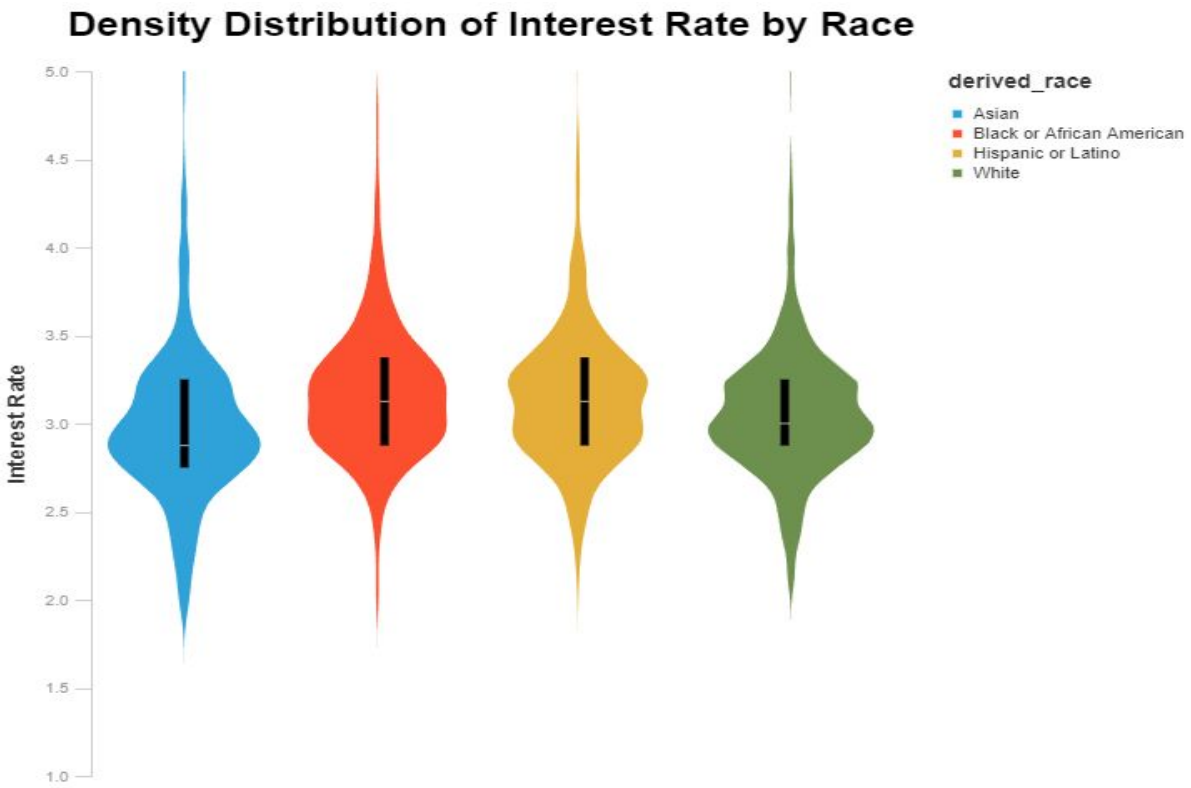
Concerns with our Results

The P-Value for Asian and Black were higher than expected, with P-Values 0.099 and 0.123, respectively. At the 5% significance level, the effects of being **Asian or Black are not statistically different than 0**, meaning that there is a reasonable chance there are no effect in being Asian or Black. The Hispanic dummy variable on the other hand, had a P-Value that was statistically significant at the 5% significance level, giving us evidence to believe that **Hispanic applicants have a higher probability of being denied a loan than White applicants.**

We believe that the reason our P-Values was high for Asian and Black was due to having less data for these two groups once we filtered our dataset to cover only Los Angeles.

Is Interest Rate Higher for Certain Racial Groups?

Denying loans to creditworthy minorities is not the only form of loan discrimination. We also considered the possibility of receiving unfair loans, or loans with higher interest rates. Here we took a step back from our Los Angeles data set and jumped into our data set covering all loans in the US. We then conducted **hypothesis tests** to see if the mean interest rates of our minority groups were the same as the White group.



Our results show that the average interest rate for **Black** was **0.09% higher** than White and the average interest rate for **Hispanic** was **0.11% higher** than White. **Asian** had **0.09% lower** average interest rate than White, signifying that they receive a more favorable interest rate than White.

Summary Statistics for Interest Rate by Race in the US

Race	Asian	Black	Hispanic	White
Mean Interest	2.98	3.16	3.18	3.07
3rd Quartile	3.25	3.38	3.38	3.25
Median Interest	2.88	3.13	3.13	3.00
1st Quartile	2.75	2.88	2.88	2.88
T - Statistics	37.40	-32.12	-51.11	NA
P - Value	0.00	0.00	0.00	NA

Although the magnitude of these differences may not seem large, when you consider the fact that most loan principal amounts are of hundreds of thousands of dollars, a 0.1% difference can equate to a significant dollar amount.

For our hypothetical test, we performed individual two-tailed t-test for each minority group in which we compared the average interest rate with that of the White group's. Our null and alternative hypothesis is therefore:

Null: Average Interest Rate of Minority group = Average Interest Rate of White group

Alt: Average Interest Rate of Minority group ≠ Average Interest Rate of White group

Because our P-Values are less than 1%, the difference between the minority groups and white's interest rate is significant at the 1% level. **We reject the null hypothesis** that the average interest rate of any minority group's interest rate is the same as the average interest rate for the White group.

We also calculated the Pearson Correlation between HRS (qualitative measure of Redlining) and interest rate, Correlation between HRS and interest rate was very low with Asian's interest rate (0.05), Black (0), Hispanic (0.05), White (0.04). Therefore, we could not find a strong correlation between HRS and interest by race.

Is Interest Rate Higher for Certain Racial Groups?

To further observe interest rate disparities, **we ran OLS regressions** in which we regressed interest rate on a dummy variable representing one of our three minority groups, while **controlling for debt to income ratio.** Unfortunately, our r**egression model had extremely poor R-squared** score due to our variables having low correlation with interest rate.

To accommodate this, we ran a **causal model test** where we matched applicants from our minority groups with applicants from the White group based on loan amount and income to compute the Average Treatment Effect. The causal model test controls for our **independent variables (income and loan amount)** and uses a **treatment variable (Black, Hispanic, or Asian)** to compute differences between the treated group and the controlled group (White).. Our results showed:

- The average treatment effect of being Black was a 0.09% higher interest rate.
- The average treatment effect of being Black was a 0.12% higher interest rate.
- The average treatment effect of being Black was a 0.1% lower interest rate.

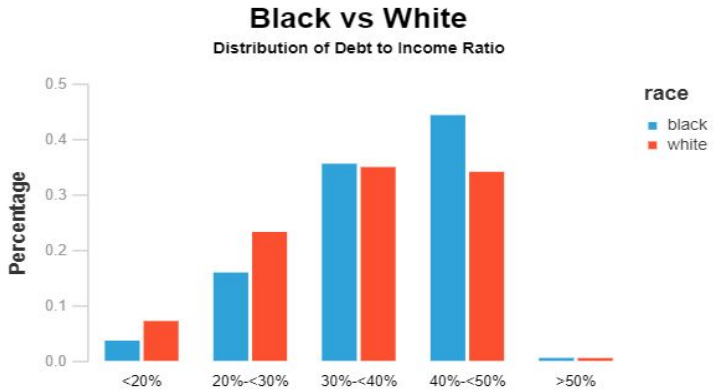
Overall, the interest rate differences between our minority groups and the White group agreed with each other, with Hispanic and Black group having about 0.1% higher interest rate and Asian group having 0.1% lower interest rate.

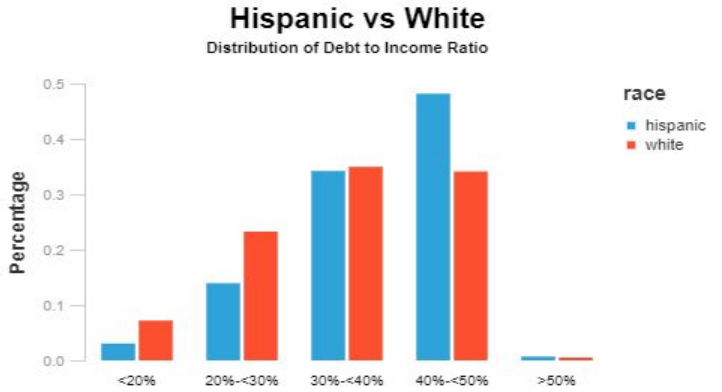
Is Interest Rate Discrepancy attributable to Discrimination?

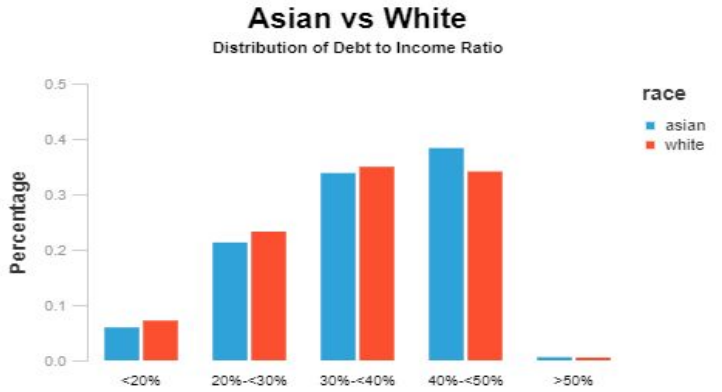
When we considered the possibilities to **why interest rate would be different** across all of our race groups, we first considered the gap in wealth. As illustrated in our Historical Redlining analysis, a **large portion of minorities occupy low grade areas** and the low grades themselves symbolize poverty. Perhaps the **interest rate and loan approval rate discrepancies are attributable to poverty** and therefore, risk. To analyze financial stability in our groups, we looked at the debt to income ratio computed by the lending institutions. Debt to income ratio is a good measure of financial stability because:

- A high debt to income ratio means more debt to be paid or less income, representing higher risk.
- A low debt to income ratio means less debt to be paid or more income, representing lower risk.

For every minority group, we plotted the **distribution of debt to income ratio** with the White group to observe differences. Our results showed that **minority groups have higher frequency in the 40-50% debt to income ratio bins** than the White group. This suggests that **loan institutions are accepting higher risk from minorities** than White applicants. The **higher risk factor from lending** to minorities **would explain the discrepancies in interest rates** and approval rates between groups.







Conclusion

Concluding Summary

When we started out on this project, we were excited to see whether loan discrimination existed in to today's society and the effects it had on minorities. Through our analysis of comparing mortgage loans from Asians, Hispanics, and Blacks to Whites, **we were able to identify differences in interest rates and approval rates for different racial groups**. By analyzing mortgage applications from 2021, we learned that Hispanic and Black applicants have roughly 0.10% higher interest rate than White applicants. In addition, Hispanic and Black applicants have a approval rate that is lower than White applicants by 7% and 11%, respectively. Depending on the area in which loans were applied to, these minority groups may experience an even lower approval rate based on the loan property's redlining score. To our surprise, Asian applicants had an approval rate that is identical to White applicants and an average interest rate that was 0.09% lower, inferring an upper hand over White applicants.

Furthermore, using our computed historical redlining score, we compared demographics of **Los Angeles County** by binned these scores and observed **significant differences in the percentage of minorities and vacancy** between different grades. We noticed signs of overcrowded homes and large proportion of vacant homes, reasons to believe loan discrimincation may still prevail,

However, having conducted our analysis through regression models, hypothesis testing, and statistical analysis, **we could not conclude the existence of discrimination nor were we able to find relationships between redlining scores and loan approval rates or interest rates.**

In addition, having analyzed mortgage loans in the US as a country, we cannot conclude that discrimination exist as we believe the discrepancies are attributed to differences in wealth and other unobservable variables like credit score.

Suggestions & Considerations

An omitted variable our team would have liked to observe was applicant's credit score. Credit score was not provided in our datasets as it would be a violation privacy. We believe credit score could be a strong determinant of interest rate and approval rate. Also, we would like to further explore the relationship between income, living conditions in redlined areas.

Our team would have also liked to observe differences of interest rate and approval rates across multiple years as a time series; however, due to the sheer size of the mortgage data sets, it would be extremely computationally intensive.

Our team could have further filtered income for better comparisons between groups but we feared it would filter out mortgage applications attributable to areas of low grade/historical redlining scores, ultimately introducing selectional bias to our analysis.

Due to changes in county lines from 1930 to 2022, counties today include multiple counties in 1930, therefore, have multiple redline grades. This may slightly hinder our Historical Redline Score estimates.

References / Citations and Statement of work

Sources

- **National Loan application records (LAR) 2021.** https://ffiec.cfpb.gov/data-publication/dynamic-national-loan-level-dataset/2021
- **Home Owner Loan Corporation (HOLC).** https://data.diversitydatakids.org/dataset/holc_census_tracts-home-owner-loan-corporation--holc--neighborhood-grades-for-us-census-tracts/resource/402d2eb0-c096-48d3-bf76-4af66f80953d
- **CensusData 1.15.post1.** https://pypi.org/project/CensusData/
- **County Code Geolocation.**
- https://github.com/kjhealy/fips-codes/blob/master/state_and_county_fips_master.csv
- https://github.com/btskinner/spatial/blob/master/data/county_centers.csv

Citations

- **Identifying Lending Disparities in mortgage.** https://revealnews.org/article/how-we-identified-lending-disparities-in-federal-mortgage-data/

- **Mapping Inequality:** https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58
- **Redlining Score.** https://ncrc.org/redlining-score/

Resources:

- Zoom: Meetings. (https://zoom.us/)
- Deepnote: Data Notebook. (deepnote.com)
- Slack: Collaboration (https://slack.com/)
- Python Libraries: PySpark: Data processing, Altair: Visualizations, SciPy: Statistics., GeoPandas: Geospatial data manipulation, CensusData: API to www. census.org

Statement of Work.

- **Kenny Tang**
  - Approval rates.
  - Income debt to ratio
  - Hypothesis testing.
  - Ordinary Least Square regression
  - Average treatment test.
  - Logistic regression
  - Report Writing.

- **Sara Haptonstall.**
  - Historic Redlining Score
  - Census demographics
  - Correlations (Interest rate, approval percentage, loan to value, etc.,)
  - Maps
  - Report design.
  - Pyspark environment

- **Nick Wu**
  - Provided input on: EDA, simple regression analysis, data cleaning and feature selection.
  - Performed Student's T-tests on several hypotheses and created exploratory and summary visualizations.