WINTER DATA CHALLENGE

INVESTIGATING BIAS IN MORTGAGE LENDING

BY

ANKUR JOSHI

SRINIVAS V PARIMI

Executive Summary

The objective of this project is to analyse and conclude if there exists any bias in mortgage lending based on different factors such as Race, Gender etc. Our analysis and conclusion bases upon the HMDA data of the state of New York and Texas for the consecutive years 2015 and 2016. This information is provided by Consumer Financial Protection Bureau, which is publically available in official website.

Through this project, we tried to investigate if there exists a racial or gender bias in loan approval process, and how this bias varies upon other pivotal factors such as Geography, time and various Income levels. The parameter that we used to measure this bias is the approval rate i.e. the % of total loan applications that get approved for a given group.

Data:

The HMDA dataset consisted of 0.9 MM and 2.4 MM records for New York and Texas states respectively and each record having 47 attributes.

Analysis:

The analysis of this data challenge rests primarily on the hypothesis proposed and how the data supports or refutes this hypothesis.

Out of total 47 variables in the datsaset, we have included following variables for our analysis:

Columns Name	Information	Missing Values
hud_median_family_income	***Median Family Income of the region***	***Yes (0.7%), Missing values were dropped from the analysis where these variables were used***
applicant_income_000s	***Applicant's Income***	***Yes(15%), Missing values were dropped from the analysis where these variables were used***
state_name	***State***	***None***
applicant_race_name_1	***Race***	***None***
applicant_sex_name	***Gender***	***None***
as_of_year	***Year of application***	***None***
action_taken_name	***Loan Status***	***None***

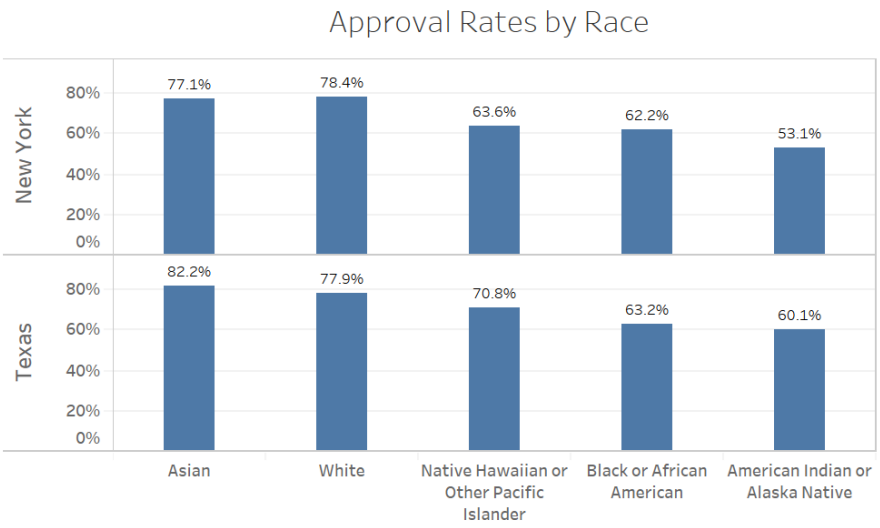
We compared loan approval rates for different groups to measure the bias in the system. Table below shows the action taken categories included in the analysis and their application status.

Action Taken	Included in Analysis	Application Status
Application approved but not accepted',	Yes	Approved
Application denied by financial institution	Yes	Rejected
Application withdrawn by applicant	No	Not Included
File closed for incompleteness	No	Not Included
Loan originated	Yes	Approved
Loan purchased by the institution	No	Not Included
Preapproval request approved but not accepted	No	Not Included
Preapproval request denied by financial institution	No	Not Included

The loan is termed as approved if the action taken is either 'Application approved but not accepted' or 'Loan originated'. The approval rate is calculated as the proportion of approved loan applications in the total number of applications: Approved/ (Approved + rejected).

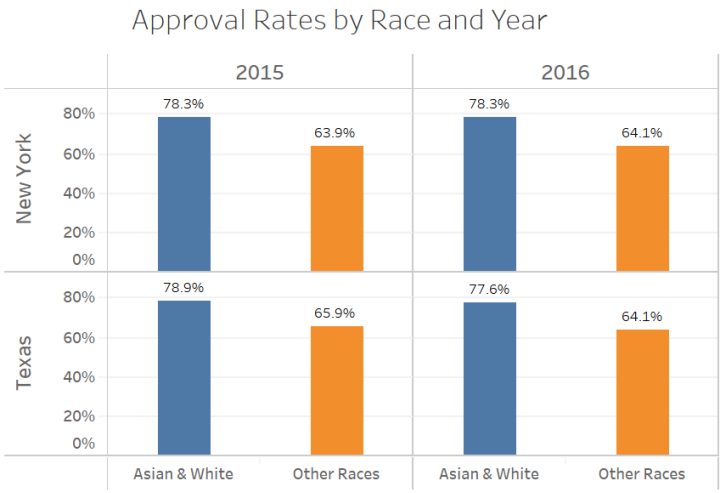
Hypothesis 1: There exists a racial bias in lending. Our intuition was that there is a bias against African Americans and other minority races as compared to Whites.

Chart below shows the approval rates for different races:



We can clearly see that the approval rates are significantly higher for Whites and Asians as compared to other 3 races. Even though, the Asians are a minority race, they enjoy a comparable approval rates to that of the majority White race, which means the bias in the lending process can't be classified into a dichotomy of minority Vs majority but it's more a race specific phenomena. Therefore, for the subsequent drill down analysis, we grouped together Whites and Asians as 'White or Asian' group and remaining three races as 'Other races'.

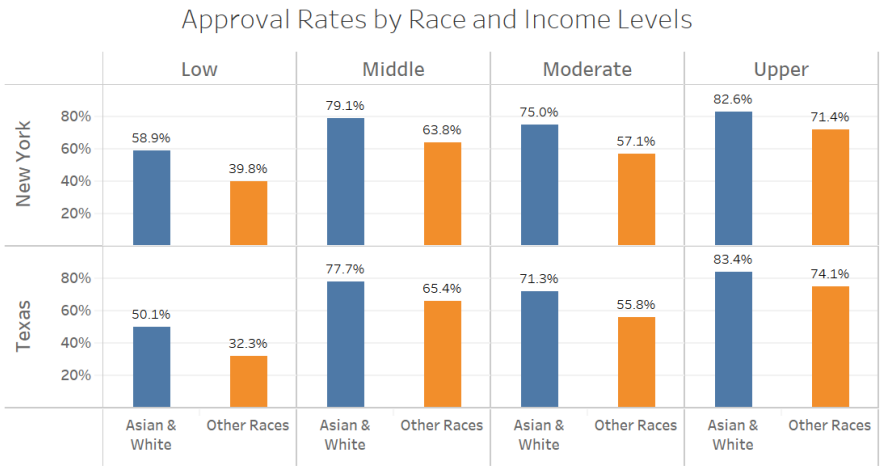
Next, we wanted to investigate how this difference in approval rates is changing over time. Chart below shows the approval rates for the two race groups in year 2015 and 2016.



It seems there is no significant change in bias from one year to another as far as the difference in approval rates are concerned. The difference remains constant from 2015 to 2016.

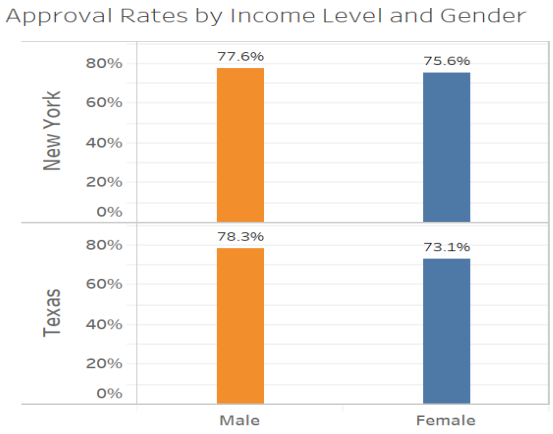
Hypothesis 2: The racial bias is higher in low-income applicants as compared to higher income ones:

This analysis uses definitions of low, moderate, middle, and upper income as proposed by the Community Reinvestment Act (CRA) regulations. In this system, the income of applicants is compared with median family income of the local area to classify them into various groups. If applicant income is less than 50% of HUD family income then it is considered low income; between 50% and 80% is moderate income; between 80% and 120% is middle income and 120% onwards is upper income level. Some records did not have data for HUD Median Income or applicant's income. We simply dropped those records for this particular analysis as replacing them with any other values would not have added any value to this analysis. We found out the approval rates for the two race groups within each income group as shown in chart below:



The chart above clearly depicts how the bias against 'other races' exists across all income groups and it gets worse for lower income groups. The difference in approval rates of the two races is higher in low income groups as compared to Upper income groups.

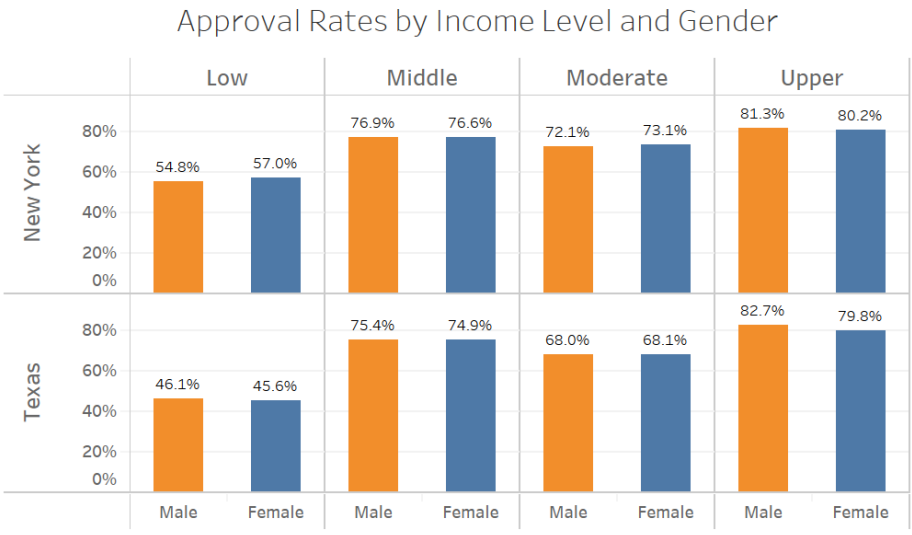
Hypothesis 3: There exists a gender bias in lending. Our intuition was that there is a bias against females in loan approval process.



Here we have excluded records that had gender reported as either "Not Applicable" or "Information not provided".

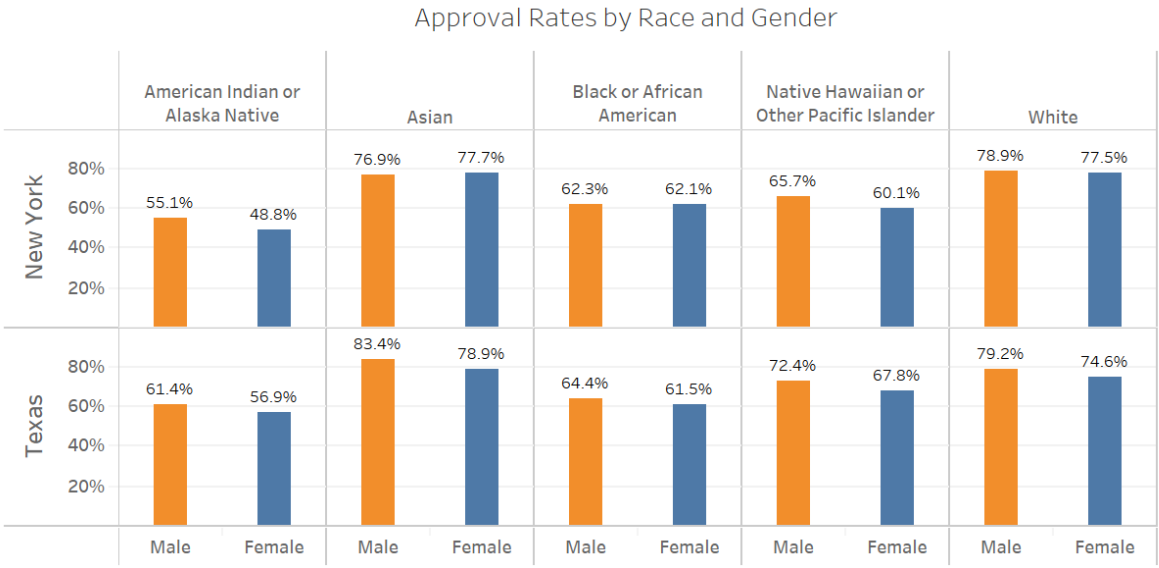
We can deduce from the chart that there is some gender bias in lending. At an overall level, females tend to have lower approval rates than males in both Texas and New York states.

Next, we wanted to investigate if this bias against females is consistent across all income levels:



We can see that the gender bias is not consistent across income levels. In fact, females enjoy a higher approval rates in low-income group in New York and moderate Income group in both Texas and New York states. Therefore, the gender bias is not uniform across income levels.

Next, we wanted to investigate if this bias against females is consistent across all races:



We observe that the gender bias against females is especially higher for American Indian or Alaska Natives & Native Hawaiian or Pacific Islanders.

Conclusion:

Here is the summary of our findings:

1. There is a clear bias against Non-White and Non-Asian races as far as the loan approval rates are concerned, and this bias is consistent across Texas and New York states.
2. Bias against these races gets accentuated for lower income groups.
3. At an overall level, there is some bias against females in both the states. This bias is not as stark as it is against some races but it exists.
4. This bias against females is especially worse in minority races like American Indian or Alaska Natives & Native Hawaiian or Pacific Islanders.

Final Comments:

Although the difference in approval rate gives an approximate indication of a bias in the lending process, to make this analysis more robust, we need to further investigate into the reasons for approval or rejection of loan applications, which can depend on several factors like the credit history, criminal record, education level etc. of the applicant. It is worth mentioning, that there does exist a variable named "denial_reason_name_1" in the HMDA dataset, which is supposed to have the information regarding the reason for rejection of an application, but for some reason, this attribute is empty for majority of the records. Hence, we have not delved into those aspects.

As a next step, we can also deep dive into the approval rates for pre-approval applications, which are excluded from our current analysis, and do a similar investigation to figure out if there is any racial or gender bias in the acceptance of pre-approval requests.