Revealing LendingClub Investor Racial Bias

Team 24

1 Executive Summary

LendingClub was a peer-to-peer lending platform allowing independent investors to accept loan requests from users. The service received heavy scrutiny for supposedly enabling lenders to discriminate against borrowers by race. In this paper, we hope to investigate allegations of racial bias in lending and analyze potential indicating factors by employing demographic data from the U.S. census, isolating the influence of different factors, and performing statistical testing. Our primary inquiry question throughout this paper is:

Can we identify a racial bias between African Americans and Caucasian in percent of accepted applications by LendingClub?

In order to show that the race of a borrower was a deciding factor in the approval of a loan, we consider the complex interplay of variables including poverty, debt-to-income ratio, and employment length verifying that even when controlling for all of them, the borrower's race is still associated with a higher loan rejection rate. We perform regression on subsets of the data under different poverty, and debt-to-income ratio conditions to qualitatively illustrate that the relation holds in all situations, and were thus able to demonstrate statistically significant correlation between the percentage of African Americans in a given region and the loan rejection rate.

Finally, we introduce a more nuanced discussion on the perceived race of the borrower and explore how this spectrum of identities effects lending decisions. We deviate from a simplified binary classification of black and white by exploring various different identities as reported in the US census. We conclude that multi-ethnic individuals who can be perceived as Caucasian are less likely to be discriminated against than those who solely identify as African American.

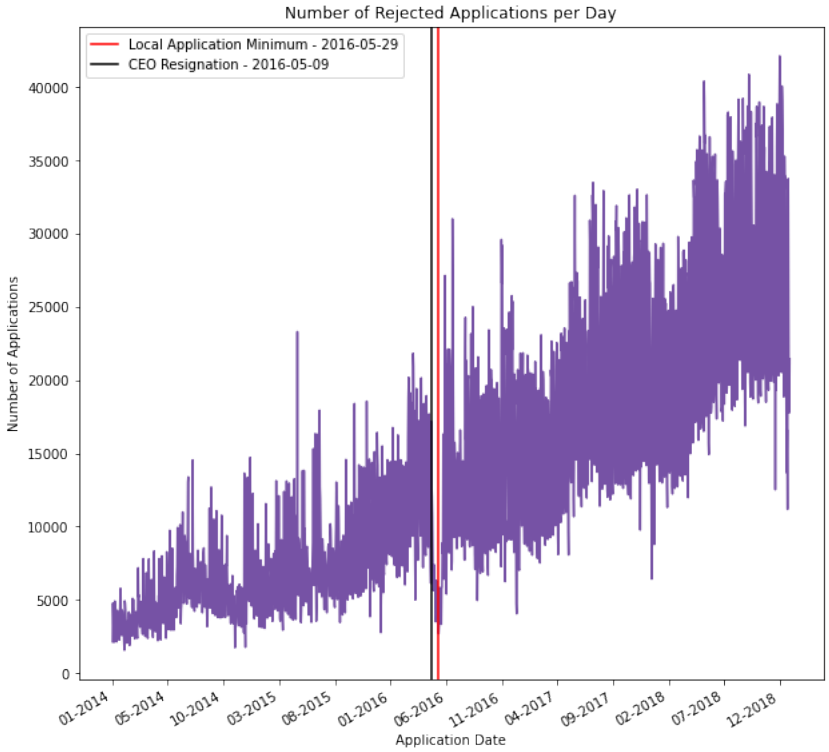
2 Data Preparation

Two data sets from LendingClub are presented: the accepted loan applications data set from 2014 to 2018, and the rejected loan applications data set, also from 2014 to 2018.

2.1 Preliminary Observation

An initial observation is that the accepted data set holds more variety in terms of applicant information, ranging from precise issue date to loan descriptions, whereas the data

presented in the rejected data set is more scarce. Therefore, we first analyze these two data sets separately, focusing on observing time series results relating to the CEO resignation in 2016. For example, the following plot of "Number of Rejected Applications" with respect to "Application Date" shows a pause for processing applications before and after the resignation.



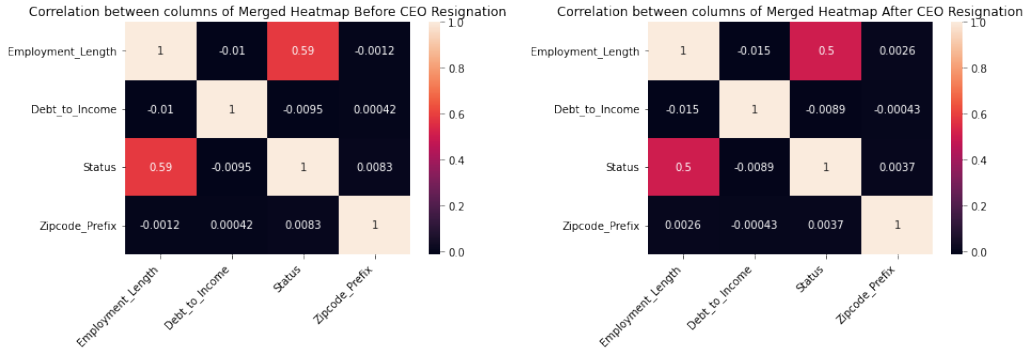
However, we do not notice a clear increasing / declining trend in the number of accepted / rejected applications.

In order to compare and contrast the two data sets alongside each other, we then decide to merge these two data sets to obtain information such as the percent of accepted applications with respect to a specific feature. Observe that there are overlaps of data type between the two data sets: We are able to extract

- **Zip Code information**
- **Loan Title**
- **Employment Length**
- **Debt to Income Ratio**
- **Policy Code**
- **Application Date**

from both data sets.

Focusing on the heat maps for numerical columns **Employment Length**, **Debt to Income Ratio**, and **Status**, we again don't notice significant changes in the correlation matrix with respect to the CEO resignation date.

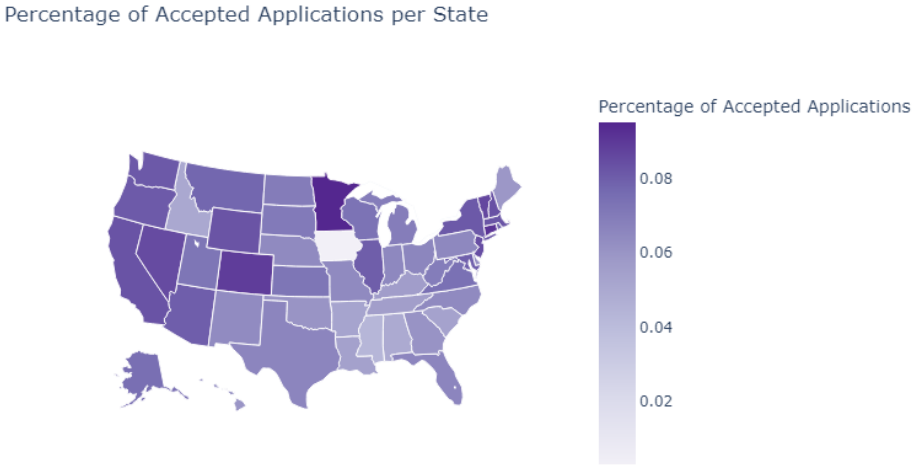


This motivates us to explore other directions focusing on specific information regarding the applicants, such as certain underlying biases. These overlapping data types provide excellent information for us to investigate potential bias that is rooted in LendingClub's business decision process.

2.2 Examining Geographical Information

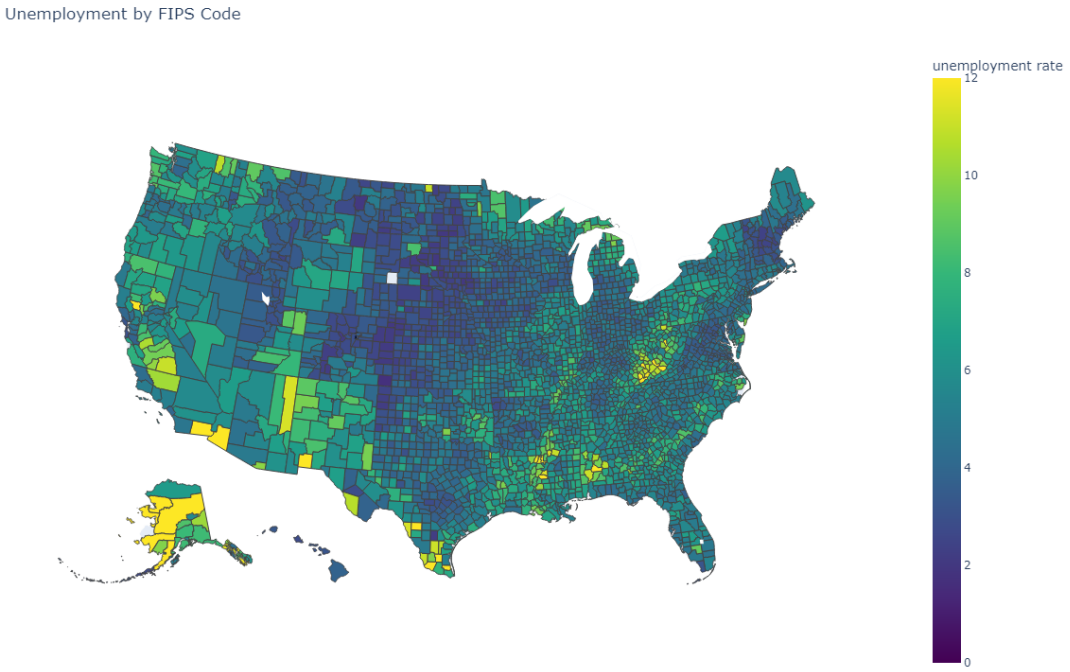
Through Exploratory Data Analysis with regards to different columns of applicant information, we found the **Zip Code information** to be of most interest to our investigation. Zip code information not only entails a person's geographical location, but also can be the basis of inference from the U.S. census data for many other factors, such as wealth, ethnic information, or social status.

By plotting preliminary **Accepted Application Percentage** versus state by merging the Accepted and Rejected datasets, we see the following output.



From visual inspection, we see a gradient of lower accepted applications towards the south-east region of the United States. Due to historical socioeconomic factors primary regarding poverty rates, we begin to explore a potential correlation between the chance an application was accepted versus socioeconomic information.

As employment length is correlated with the status of an application as noted in **Preliminary Observation**, we explore the unemployment trends in the United States. To do so, we plotted unemployment by FIPS region to visually identify a correlation between the status of an application and unemployment.



2.3 Integrating Our Data set

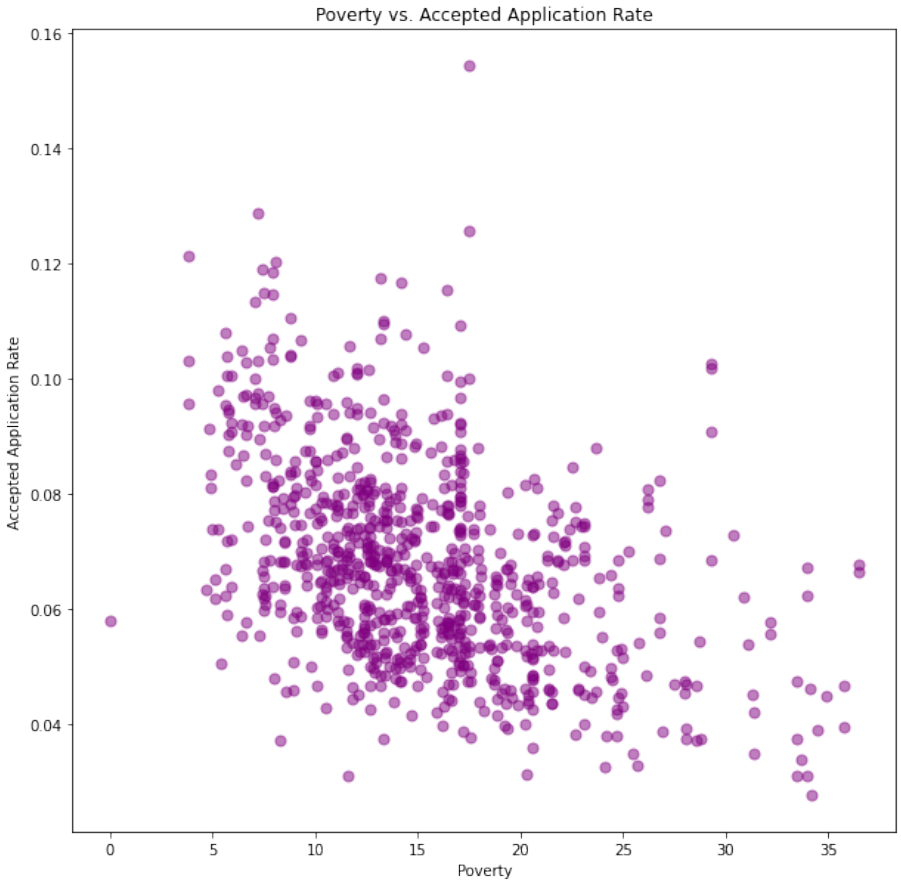
We acquire Census data to integrate with the original geographical information from the datasets. We propose to use the demographics data on poverty to assist our investigation.

2.3.1 Zip code Prefix to FIPS

While the provided data sets contained a zip code feature with the first three digits of the application's zip code, accessible poverty information is broken down by FIPS code, or a Federal Information Processing Standard that breaks down US counties into a five digit code [3]. Each zip code prefix in the data set was matched to a corresponding FIPS code via an external dataframe.

2.3.2 Poverty from FIPS

With census-provided FIPS poverty rates [2] matched to the original application dataframes, application data is grouped by zip code prefix for further analysis. This grouping consists of average poverty rate, average employment length, average application acceptance rate, and average debt to income for the zip code.



The scatter plot above indicates a negative correlation between poverty and the likelihood that an application was accepted. In other words, in zip codes with a higher poverty rate, applications from that region were less likely to be accepted. This preliminary qualitative data supports our initial exploration of our hypothesis.

2.3.3 Demographic Data

Lastly, to explore potential demographic biases within the application data sets, demographic information is sourced from further census data. By matching application zip code groups to the ZCTA5 [1] standard within the census demographic data, averages of racial information are included in each zip code grouping.

3 Demographic Biases in Loan Acceptance

With over 900 unique valid zip code groupings of varying racial identity from across the United States, we make the decision to focus insight into potential demographic biases within the data towards African American.

3.1 Encapsulating Loan Acceptance

To understand what metrics are critical in determining the predictability of a loan's status from race, OLS regression is conducted upon both racial categories by zip code corresponding to African American and Caucasian. The features, Poverty, Employment Length, Percent African American by Zipcode, and Debt to Income are chosen based upon the correlation matrix outlined in **Data Preparation**. The regression yields an R^2 of 0.832 and t-test p values of approximately 0 for each feature. Thus, we can reject the null hypothesis and accept the possibility that each of these features play a role in determining the acceptance chance of a loan for African Americans in the sample.

An identical OLS regression is performed upon Employment Length, Poverty, Percent Caucasian by Zipcode, and Debt to Income yielding an R^2 of 0.818. Similarly to the above findings, Employment Length and Debt to Income correlate quite highly with predicting the chance that a loan application was accepted. Percent Caucasian by Zipcode, however, yielded a t-test p value of 0.6, failing to reject the null hypothesis.

Thus, from our preliminary encapsulation of loan acceptance prediction, we note that a zip code's percent of Caucasian population is seemingly uncorrelated to the outcome of the loan, while the percent African American population of a zip code does.

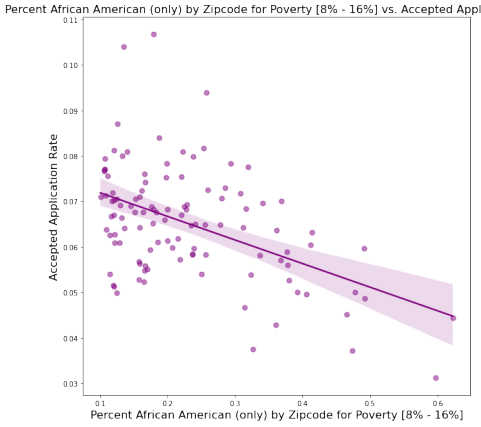
3.2 Demographic Isolation for Revealing Biases

While Employment Length, Poverty, and Debt to Income are correlated with the likelihood of a loan being accepted, per our hypothesis we must isolate each correlated factor to determine the influence of racial bias in the acceptance of a loan.

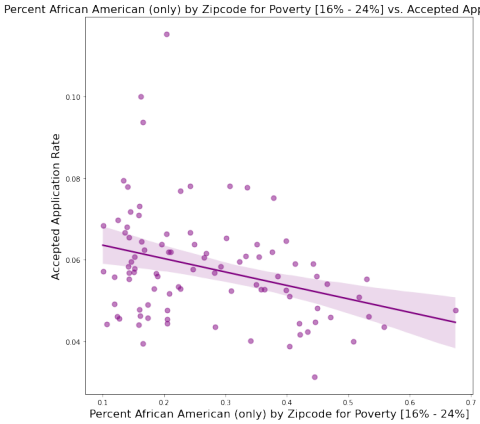
First, we note that while Employment Length is correlated with status, on average, the spread and distribution by zip code, irrespective of race, is incredibly tight and thus per zip code grouping does not play a role in exploring our hypothesis. Debt to Income and Poverty, however, are correlated with race and must thus be isolated and examined.

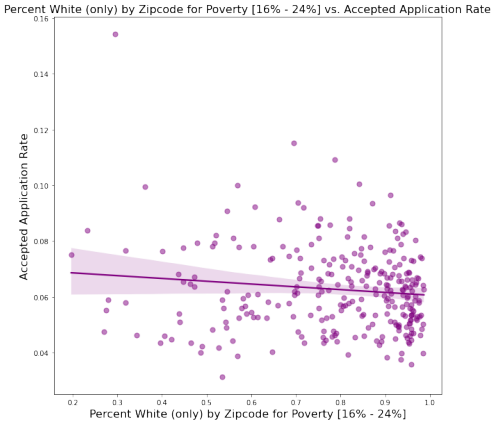
We attempt to isolate the poverty feature from the racial statistic. Each racial demographic group is isolated into **poverty bands** of 8% to determine inter-band correlation. In other words, by isolating for poverty, we explored if the trends shown prior still exist. Scatter plots with regression are isolated by African American percentage (left) and White/Caucasian (right). (We leave the 0% - 8% and 24% - 40% range to Appendix A, as the majority of the data falls within the 8% - 24% range.)

- Poverty Band 8% - 16%



- Poverty Band 16% - 24%





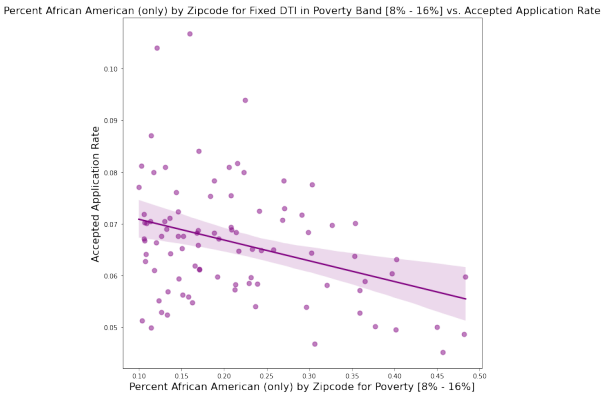
Race	Coefficient	Standard Error	t-statistic	p-value
African American	-0.0169	0.004	-4.548	0.000
Caucasian	-0.0017	0.003	-0.518	0.606

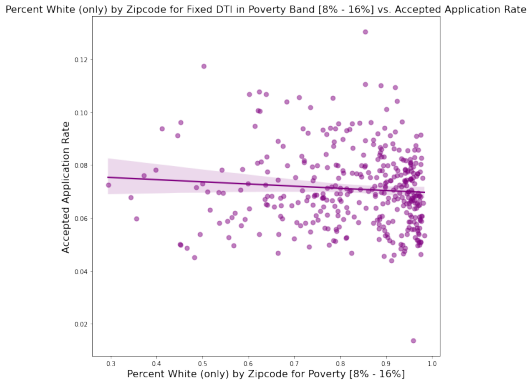
Table 1: OLS Analysis Results for African American / Caucasian Applications

Additionally, we see that in the OLS regression that there is high negative correlation between higher African American population in a zip code and the status of a loan being accepted in all poverty bands indicated by p-values near zero. For all the African American segmented plots, we observe a p-value of 0.000, indicating that the correlation between the Accepted Application Rate and Percent of African American population by zip code is statistically significant. There is little to no correlation in the Caucasian dataframes as indicated by p-values above 0.5 and low across-the-board R^2 values.

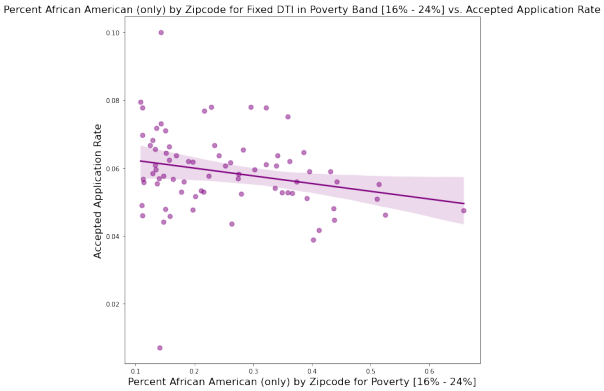
Next, as this data is not isolated for Debt to Income, the existing scatter plots were further segmented by **bands of Debt to Income (DTI)**. The same bands of the most populous Debt to Income ratio is selected, but the trends are similar in all 20 plots. There is insufficient data for Debt to Income ratios in the 32%+ poverty bands. (We leave the 0% - 8% and 24% - 32% range to Appendix B.)

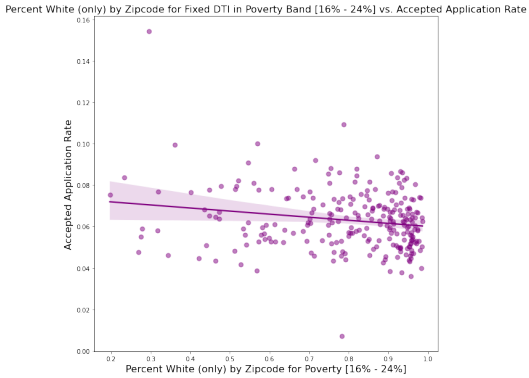
- Poverty Band (With fixed DTI) 8% - 16%





- Poverty Band (With fixed DTI) 16% - 24%





Similarly with the findings above, in the African American segmented scatter plots we find p-values for each regression that rejects the null hypothesis and states that for each poverty band, even when isolating for Debt to Income, higher population African American zip codes have a lesser likelihood of having an accepted loan application. The p-values in the Caucasian scatter plots are not statistically significant.

4 Racial Perceptions in Loan Acceptance

In the previous section we have shown that the decision to fund a loan on LendingClub has been influenced by racial identity. Furthermore, we claim that the socially perceived race of an applicant also plays a factor in determining whether a loan is funded. In this discussion, we claim that individuals less perceived as Caucasian have a lesser likelihood of having a funded loan.

4.1 Feature Construction and Analysis

In the Census-sourced data set for demographics, racial identity is split into broad categories of varying mixed-race identities. For our hypothesis we focus on analyzing racial perceptions for African American identifying individuals within the data set.

Furthermore, we establish three boundary conditions to isolate demographic groupings within racial designations.

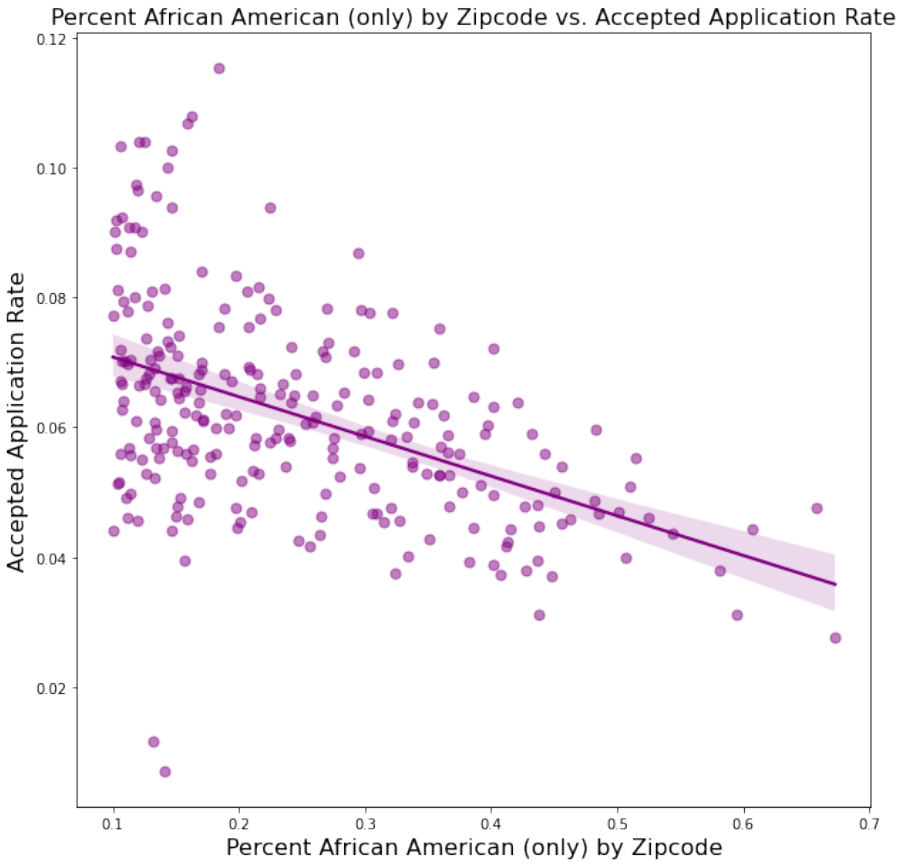
1. Single-Response Racial Identity: Individuals who report a single-race on the census form

2. Mixed-Response Racial Identity: Individuals who report any combination of African American and Caucasian

3. 6-or-More Response Racial Identity: Individuals who report African American as one of six or more components in their racial identity

From the above census demographic classifications, we focus on Single-Response Racial Identity and Mixed-Response Identity. While 6-or-More Response Racial Identity gives insight into racial perceptions, the category is too broad for rigorous exploration and will thus be excluded from the analysis.

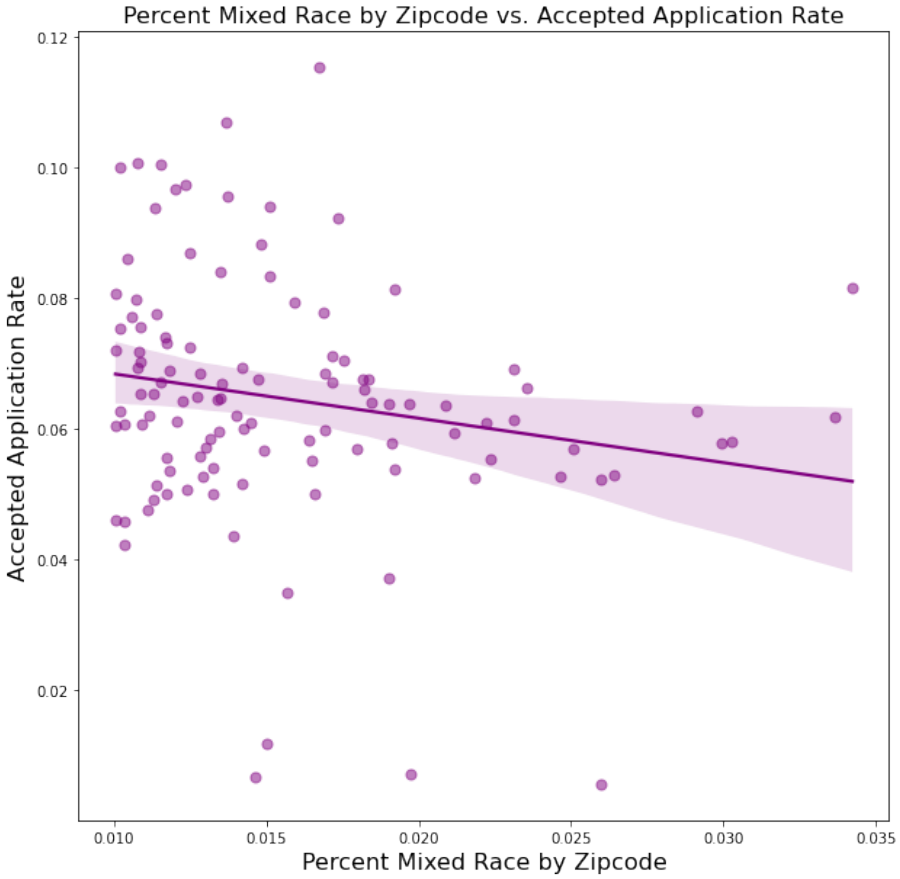
As a baseline, we plot zip codes on their racial identity for individuals who only responded as "African American" on the census data versus the zip codes' average loan application acceptance rate.



With a standard Ordinary Least Squares regression yielding an R^2 of 0.652 and $P > |t|$ of approximately 0, we reject the null hypothesis that changes in Percent African American are uncorrelated with Accepted Application Rate. We have statically significant evidence

to conclude that applications are less likely to be accepted in zip codes with a higher population of individuals who respond as African American as their sole racial identity.

We repeat the process upon the Mixed-Response Racial Identity set. With an identical regression, we yield R^2 of 0.036 and a p-value of 0.036. While the p-value of this regression allows us to reject the null hypothesis in a similar vein to the former dataset, we note that the negative correlation is much weaker. Thus, we make the qualitative conclusion that as individuals are perceived as "more white", they are most likely to be grouped with other individuals who are perceived as white. As demonstrated above with racial perceptions shaping loan acceptance, we conclude that as individuals are more likely to be interpreted as "white", their loan is more likely to be accepted.



5 Conclusion and Discussion

Throughout a deep exploration of the LendingClub datasets and U.S. census dataset, we can give an affirmative answer to our inquiry question that *statistically significant racial bias is indeed observable amongst the fulfilled loans in LendingClub application pool.*

From explorations regarding demographic data and relating the approved rate to racial information of the applicants, we are able to observe a clear contrast between correlation of population densities of African Americans and Caucasian to Approved Application Rates through graphs and OLS regression analysis. This gives us evidence to establish the claim

that racial information indeed influences one's chance of obtaining a loan on the LendingClub platform.

Furthermore, while racial identity plays a critical role in determining the outcome of a loan application, an investor's perceived interpretation of the applicant's race is also critical. In other words, if the investor is able to perceive an applicant as "African American", then their loan, on average, is less likely to be approved, even when accounting for most correlated factors to loan acceptance. Thus, we have shown that due to the nature of LendingClub, investors not only implicitly bias against African Americans in general, but also approve loans for this demographic at a lesser rate the less "white" the applicant is.

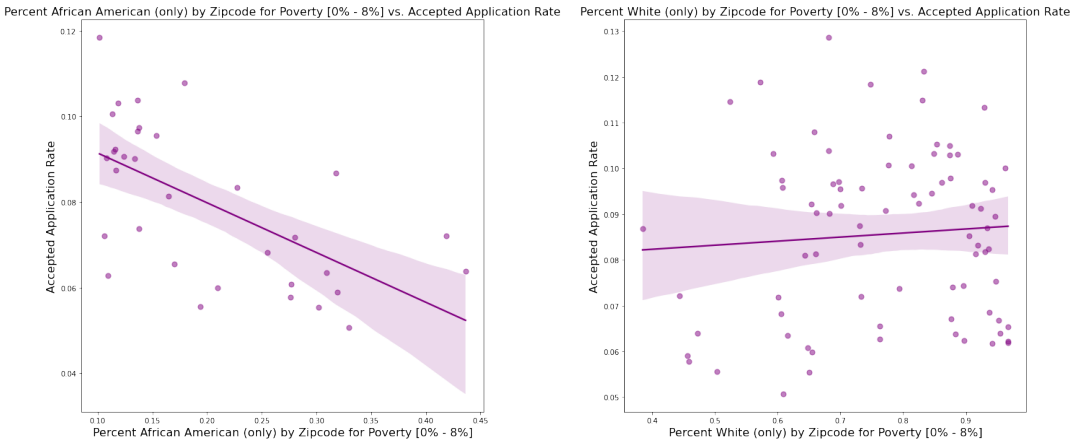
Indeed, racial bias still stands as a great challenge in our current society towards equality and equity, whether it is obtaining loans that could be a true opportunity for someone initiate a new life, or something that could concern a person's livelihood. Whether such bias is implicit or explicit, we now have learned certain tools to distinguish such bias and can work towards rectifications and change. Unfortunately, we do not have time to explore in depth of the vast amount of information presented in the LendingClub data sets, and we only utilized less than 10 percent of the columns in the provided accepted data set to draw our conclusion. However, we would be very interested to continue diving deeper into the data, perhaps running NLP analysis on columns such as EMP_TITLE and PURPOSE or generating clustering graphs to find coherent trends.

6 References

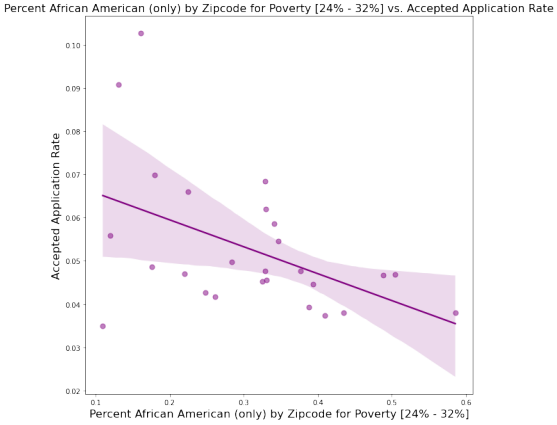
- [1] Bureau, U. S. C. (2022, May 16). ZIP code tabulation areas (zctas). Census.gov. Retrieved July 24, 2022, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html
- [2] https://github.com/Ro-Data/Ro-Census-Summaries-By-Zipcode
- [3] https://github.com/Data4Democracy/zip-code-to-county

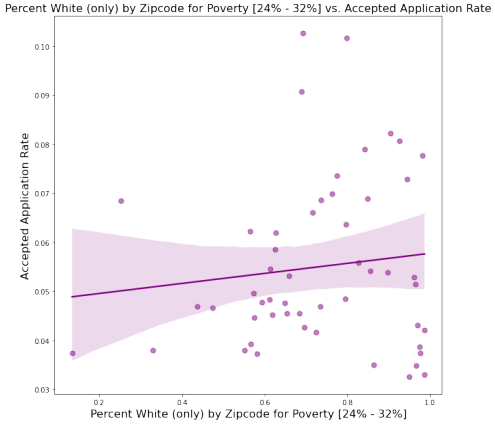
7 Appendix A

- Poverty Band 0% - 8%

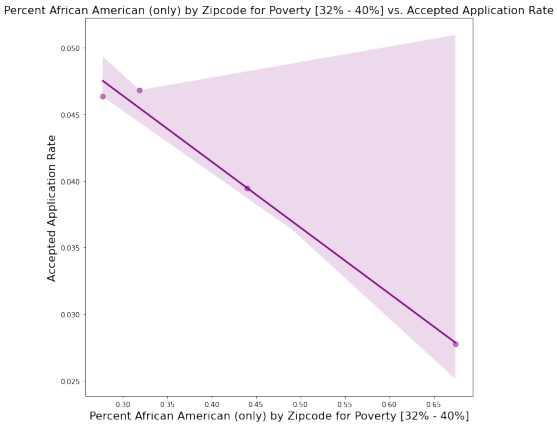


- Poverty Band 24% - 32%



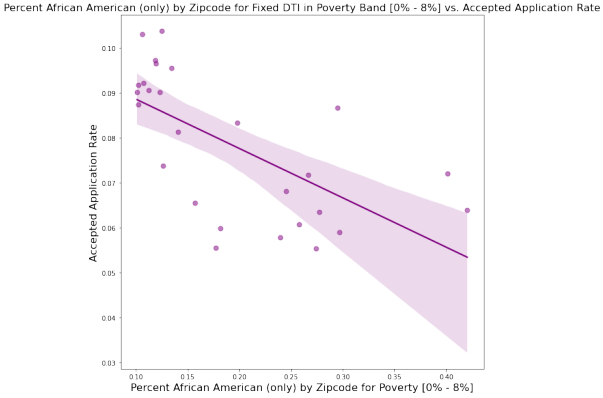


- Poverty Band 32% - 40%



8 Appendix B

- Poverty Band (With fixed DTI) 0% - 8%



- Poverty Band (With fixed DTI) 24% - 32%