ANALYSIS OF LENDINGCLUB'S GEOGRAPHICAL DISCRIMINATION OF P2P LOAN APPLICATIONS

July 24, 2022

Contents

Michelle Liu
Stephen Yin
Jeremy Lee
Samantha Lee

1 Topic Question

In the past decade, the economy has experienced significant turmoil, and credit access has become an increasingly pertinent issue for households across America. Loans are especially important as they help individuals out of tough financial situations and allow businesses to scale up and grow the economy. However, credit access is not universal across the United States. People living in rural regions of America tend to have less access to financial institutions and according to the Consumer Financial Protection Bureau (CFPB), are more likely to live in banking deserts [5]. With fewer banking options, people in rural areas may find peer-to-peer (P2P) lending, where individuals directly obtain loans from other individuals, particularly attractive.

Since its 2007 inception, LendingClub's was a pioneer in the P2P lending space. It offered accessible loans for borrowers and financial opportunities for investors, all while profiting from a small funding fee. In 2016, however, LendingClub's founder was forced to resign after being charged for fraud. Four years later, the company shifted its business practices, phasing out P2P lending and transitioning to institutional investors instead [2].

Because rural regions on average have less credit access, we wanted to analyze if in the years of LendingClub's P2P business model, people from rural areas experienced differential treatment in loans. We posed the following topic for analysis: *How does LendingClub treat different geographical regions, and has the 2016 resignation affected that?* Based on our topic, we answered two questions:

1. Can we use clustering algorithms to identify geographical discrimination by LendingClub?
2. Did LendingClub's practices change after the resignation of their founder and CEO, Renaud Laplanche, in 2016?

2 Executive Summary

In this report, we used LendingClub's loan data, provided by Citadel, and zip-code-level and state-level demographic data, provided by the US Census. We sought to answer the questions of whether LendingClub's P2P lending model discriminated against people based on geography (particularly those living in rural areas) and whether discriminatory practices changed after the resignation of the CEO in 2016. After cleaning and analyzing the data, we came to the following key findings:

1. The distribution of loans varied significantly across state borders. Loans from more rural states tended to have lower acceptance rates, lower grades, and higher interest rates.
2. While we suspected that rural proportion of a loan's state would be positively correlated with interest rates, results from a linear regression over the full set of data did not reflect that and instead found that loan grades were the strongest predictor of interest rates.

3. After clustering loans with similar grades, FICO scores, and debt-to-income ratios, we found that a particular cluster of loans exhibited a moderately strong positive relationship between rural origins and interest rates (correlation coefficient of 0.535). This cluster was characterized by individuals with high loan grades, mid-to-low FICO scores, and mid-to-low debt-to-income ratios. After some causal analysis, we determined that there was indeed a positive effect of rurality on interest rates. For individuals in other clusters, we did not find a particularly strong correlation between a state's rural population and interest rates.

4. When comparing the effect of population density of origin (a proxy for rurality) on an application's acceptance chance pre- and post-CEO resignation in 2016, we found that an increase in population density did increase an application's success rate post-resignation. The significance of the increase is debatable, but visual representations of population density's impact suggest that more sparsely populated areas (i.e. rural areas) may have been less discriminated against pre-resignation. More rigorous analysis is certainly prompted.

Equal loan access is an important part of creating equitable economic opportunity. Our results indicate that LendingClub's P2P business model did not accomplish that with respect to treating people of different geographies equally. We hope that our report sheds light on one instance of discriminatory practices in lending and encourages further studies on present-day lending data. We finally also hope for our results to bring more attention to lesser-talked-about forms of discrimination that can adversely affect vulnerable populations.

3 Technical Exposition

3.1 Data Preprocessing

With 151 columns within the accepted loan data and 8 columns within the rejected loan data, our first task in preprocessing was narrowing down the relevant columns we would run our analysis on. First, in order to make comparisons between the set of accepted loans and the set of rejected loans, we took note of the seven features that were shared between them (Figure 1 column 1). Then, when examining the remaining features in the accepted loan data, we selected ones that we thought most holistically characterized a loan (Figure 1 column 3). We dropped the columns describing delinquencies or number of open credit lines because our exploratory analysis found that many of these variables were highly correlated and could be best represented by a borrower's FICO (credit) score.

Finally, when checking for missing data, a relatively small proportion of data was missing for each column, so we decided to drop those rows. While a missing entry in columns such as *Employment Title* and *Employment Length* may be an indicator that a person is unemployed, we decided against imputing data based on that assumption. Removing missing columns ultimately led to a 3.82% reduction in data.

Selected Features			
Accepted and Rejected Loans	Missing Rows	Accepted Loans	Missing Rows
Loan / Requested Amount	0.000	Loan Amount	0.000
Application / Issue Date	0.000	Issue Date	0.000
Loan Title	0.001	Loan Title	0.000
Debt to Income Ratio	0.000	Debt to Income Ratio	0.001
Zip Code	0.000	Zip Code	0.000
State	0.000	State	0.000
Employment Length	0.033	Employment Length	0.068
		Loan Grade	0.000
		FICO Score	0.000
		Home Ownership	0.000
		Interest Rate	0.000

Figure 1: Feature selection and missing values

After cleaning missing data, we were still left with 25+ million entries of total data, so for computational efficiency, whenever regressions were run on the full data set of both accepted and rejected loans, a random sample of 1 million accepted loans and 1 million rejected loans was used. This subset was also balanced on pre- vs. post- CEO resignation.

One column that we took particular care in cleaning was *Loan Title*, as it described the purpose of each loan. We found that user inputs for this field were quite inconsistent in the words/phrases used and capitalization, so by standardizing cases and performing a keyword replacement (e.g. "credit" or "cc" to "Credit Card"), we bucketed the loan titles into seven main categories: debt consolidation, credit card, home, car, medical, business, and other. After doing this, we found that debt consolidation and credit card payoffs made up the large majority of both accepted and rejected loans (Figure 2).

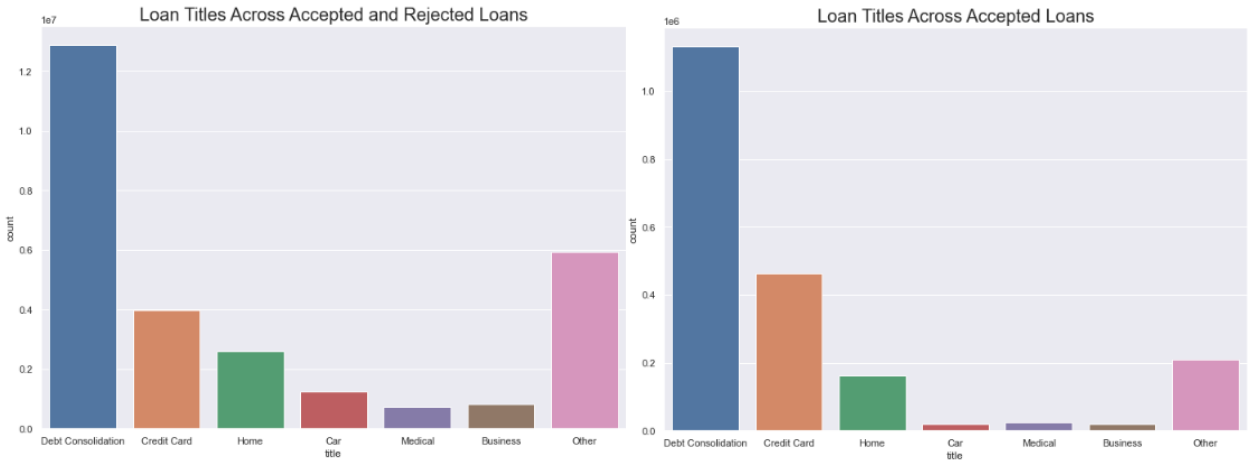


Figure 2: Loan Title Distributions

We then re-typed columns appropriately. We converted issue/application dates to standardized datetime objects and made sure formatting was consistent within each feature of the data.

3.1.1 Demographic Data from the US Census

We also brought in demographic data from the US Census, including urban-rural population numbers on both a zip-code-level and a state-level, to better analyze geographic distributions [1]. Using state-level data, we computed a column *Rural Proportion* as the number of rural households in a state divided by its total number of households.

At a zip-code-level, we used python's `uszipcode` package to retrieve a variety of demographic information including gender, race, population density, median household income, etc. In Section 3.5 of our report, we use these variables as controls when analyzing the effect of rurality on LendingClub's acceptance rate for applications.

Because the data provided did not include full 5-digit zip codes and only had 3-digit prefixes, we computed aggregate demographic information for each 3-digit prefix. For each 3-digit prefix, we calculated race, gender, and income attributes by taking weighted (by population) averages across zip codes that shared that prefix. Population density (our metric for rurality) was computed by summing the populations of each zip code sharing the same prefix and dividing the sum by the total square mileage of the zip codes.

Given only the first 3 digits of a zipcode, we aggregated the available data from all zipcodes beginning for each 3 digit zipcode present in the. While map data was also provided, run times were too long for generating a national map split by 3 digit zipcodes, hence, map data was unused for modelling and we stuck with state level visualizations.

3.2 Geographic Visualizations

To explore our question of whether LendingClub's business practices discriminated by location, we generated a number of visualizations on a state-level. First, we examined what the proportion of rural households looked like across the US to have a comparison point for the rest of our visualizations (Figure 3).

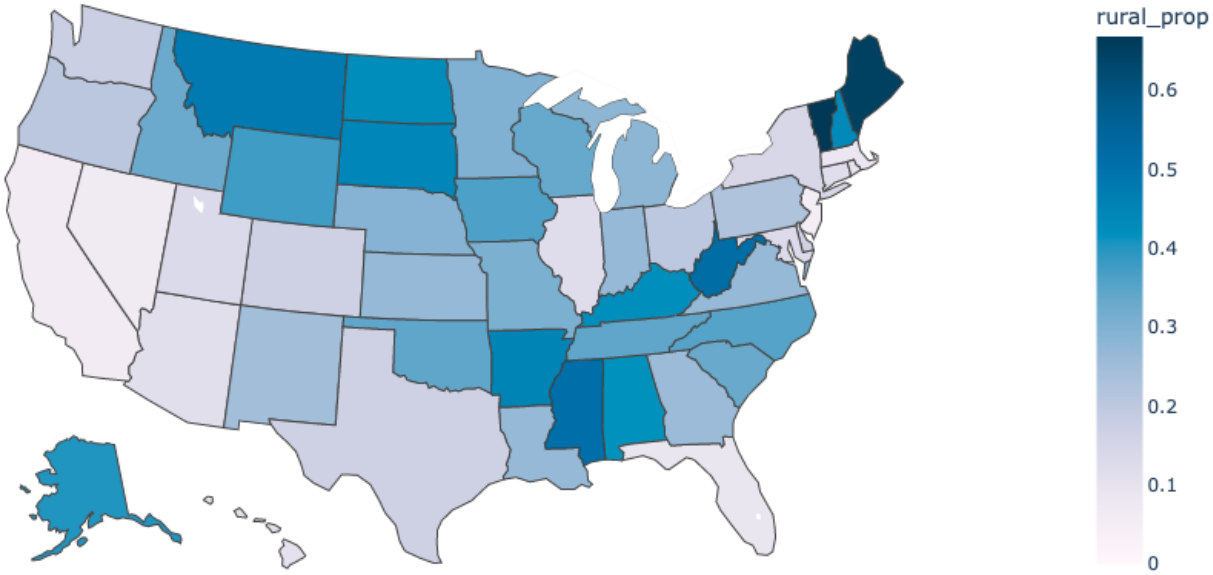


Figure 3: Rural Household Count as Proportion of State Population

The first metric we evaluated was the acceptance rate of loans, which we calculated as the number of accepted loans in a state divided by the total number of loan applications for that state. This value for all states was fairly low, since the overall number of rejected loans far outweighed the number of accepted loans in the data set. Visually, it seemed as though more urbanized states such as New York, California, New Jersey, and Nevada had the highest loan acceptance rates, while more rural states had lower rates (Figure 4).

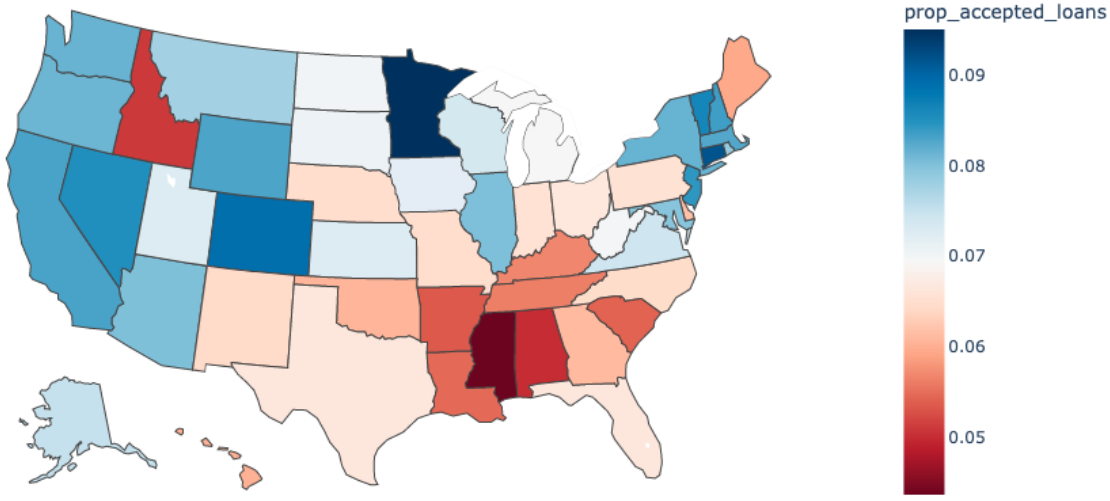
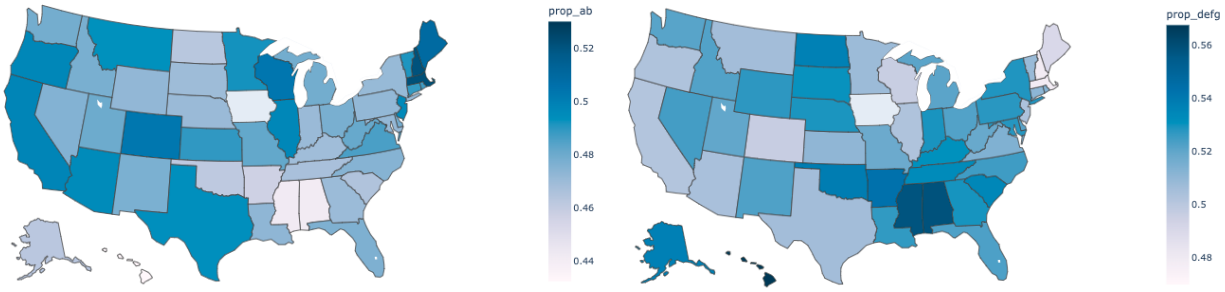


Figure 4: Loan Acceptance Rate by State

We then looked at the proportion of loans in each state that fell under particularly high (A-B) and low (D-G) loan grades (Figure 5). We found that loans were graded particularly poorly in Alabama and Mississippi. Note that Iowa was excluded from this analysis since it only had one accepted loan, and a value of 0 or 1 in the scale would have thrown off the interpretability of the choropleth maps.



(a) Proportion of Loans with High Grades

(b) Proportion of Loans with Low Grades

Figure 5: Loan Grades Across States

We finally looked at interest rates, arguably, the most important aspect of a loan, as high interest rates characterize low desirability for borrowers and high risk for lenders. Again, we observed a slight bias of higher interest rates for rural areas, particularly in Mississippi and Alabama (Figure 6).

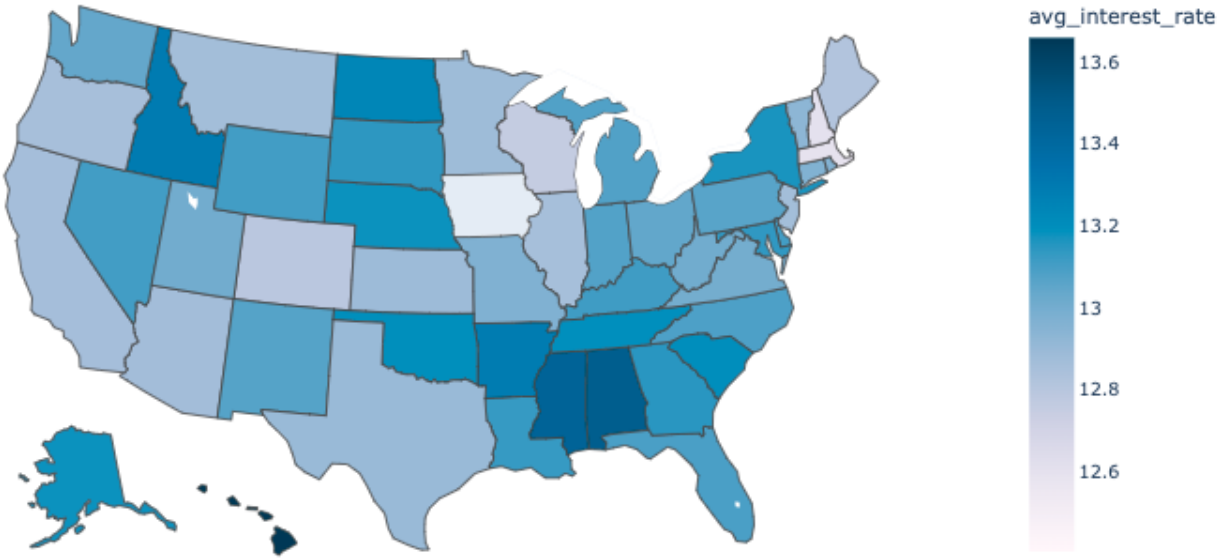


Figure 6: Average Interest Rate Across States

3.3 Descriptive Regression on Interest Rates

In order to more rigorously test for geographic discrimination, we used linear regression as our initial framework for determining which variables are strong or weak predictors of interest rates. Using an ordinary least squares regression model, we first created a baseline without involving rural population statistics by regressing interest rates against loan grade (encoded as binary indicator variables, excluding A), FICO score, and debt-to-income ratio (Figure 7).

OLS Regression Results

Dep. Variable:	int_rate	R-squared:	0.916
Model:	OLS	Adj. R-squared:	0.916
Method:	Least Squares	F-statistic:	2.755e+06
Date:	Fri, 22 Jul 2022	Prob (F-statistic):	0.00
Time:	18:04:11	Log-Likelihood:	-3.5805e+06
No. Observations:	2028241	AIC:	7.161e+06
Df Residuals:	2028232	BIC:	7.161e+06
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.3610	0.025	375.877	0.000	9.312	9.410
B	3.4027	0.003	1108.037	0.000	3.397	3.409
C	6.8657	0.003	2143.520	0.000	6.859	6.872
D	10.9518	0.004	2893.849	0.000	10.944	10.959
E	14.7758	0.005	2997.209	0.000	14.766	14.786
F	18.7472	0.008	2319.488	0.000	18.731	18.763
G	21.5231	0.014	1529.231	0.000	21.496	21.551
fico	-0.0033	3.4e-05	-96.530	0.000	-0.003	-0.003
dti	0.0042	6.8e-05	61.266	0.000	0.004	0.004

Figure 7: Interest Rate Regression on Loan Grade, FICO, DTI

Since the model had an R^2 value of 0.916, we could tell that the linear regression model fit the data well. Furthermore, from the magnitude of the coefficients, we could tell loan grade was by far the strongest predictor of interest rates.

Next, we added *Rural Proportion* of state population to our list of covariates to see if it would make a meaningful difference in our regression results. Interestingly, the R^2 value did not increase, and the magnitude of the coefficient of our new variable was negligibly small, indicating that it was not a good predictor of interest rates (Figure 8).

OLS Regression Results

Dep. Variable:	int_rate	R-squared:	0.916
Model:	OLS	Adj. R-squared:	0.916
Method:	Least Squares	F-statistic:	2.449e+06
Date:	Fri, 22 Jul 2022	Prob (F-statistic):	0.00
Time:	15:47:14	Log-Likelihood:	-3.5804e+06
No. Observations:	2028241	AIC:	7.161e+06
Df Residuals:	2028231	BIC:	7.161e+06
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.3663	0.025	375.677	0.000	9.317	9.415
B	3.4028	0.003	1108.051	0.000	3.397	3.409
C	6.8658	0.003	2143.478	0.000	6.860	6.872
D	10.9519	0.004	2893.742	0.000	10.945	10.959
E	14.7761	0.005	2997.127	0.000	14.766	14.786
F	18.7474	0.008	2319.492	0.000	18.732	18.763
G	21.5233	0.014	1529.244	0.000	21.496	21.551
fico	-0.0033	3.4e-05	-96.451	0.000	-0.003	-0.003
dti	0.0042	6.81e-05	61.426	0.000	0.004	0.004
rural_prop	-0.0396	0.009	-4.524	0.000	-0.057	-0.022

Figure 8: Interest Rate Regression on Loan Grade, FICO, DTI, Rural Proportion

Since there was a lack of strong evidence for geographic discrimination when examining the *entire* set of accepted loans, in the next section, we decided to investigate segments of our data to look for geographic discrimination within particular types of loans.

3.4 Clustering Analysis

Our goal was to segment the data set of accepted loans by clustering similar ones. By analyzing loan applications with similar characteristics but across different regions, we can better isolate for the effect of geographical bias, if there is any. To accomplish this, we used K-Means clustering on three features: loan grade, FICO score, and debt-to-income ratio. Since K-Means clustering is not operable on categorical data, we converted loan grades to ordinal form (e.g. A maps to 0, B maps to 1, ...) to preserve their ordering. We also standardized each column using the formula below such that the clustering algorithm would not be biased by variables with have higher variances than others. Prior research has demonstrated how non-standardized measurement units can produce different and less-objective clustering structures [4].

$$Z = \frac{X - \overline{X}}{s} \text{ , } \overline{X} = \text{sample mean , } s = \text{sample standard deviation}$$

We decided on the number of clusters to use by examining the inertia, or sum of squared distances from a cluster's centroid, for different numbers of clusters. Based on the elbow graph in Figure 9, we chose 5 clusters, as this was where the marginal decrease in inertia began to drop off substantially.

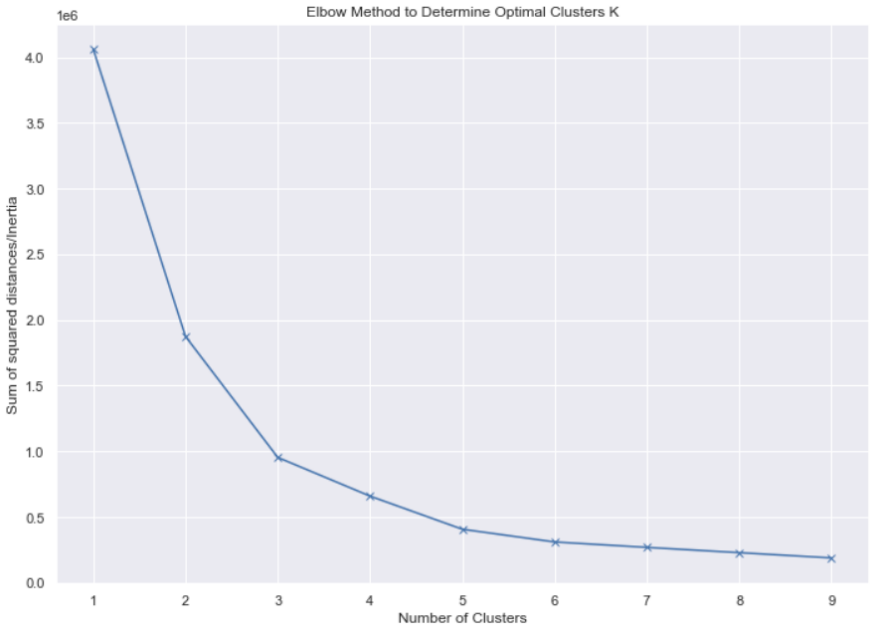


Figure 9: Elbow Graph to Determine Cluster Number

To get a better sense of what a characteristic sample looked like in each cluster, we plotted the centroids of the clusters in a 3-dimensional space (Figure 10).

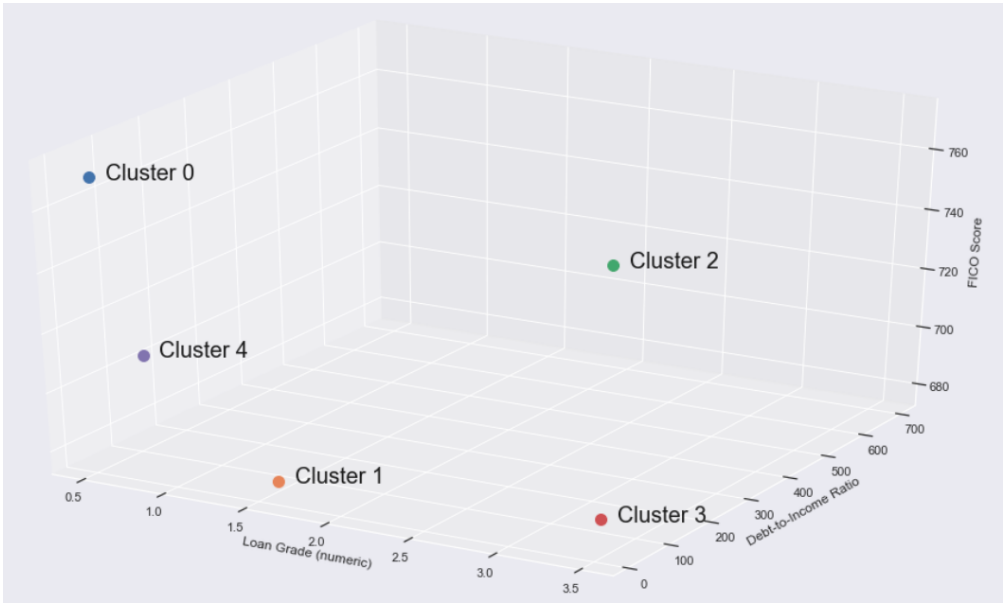
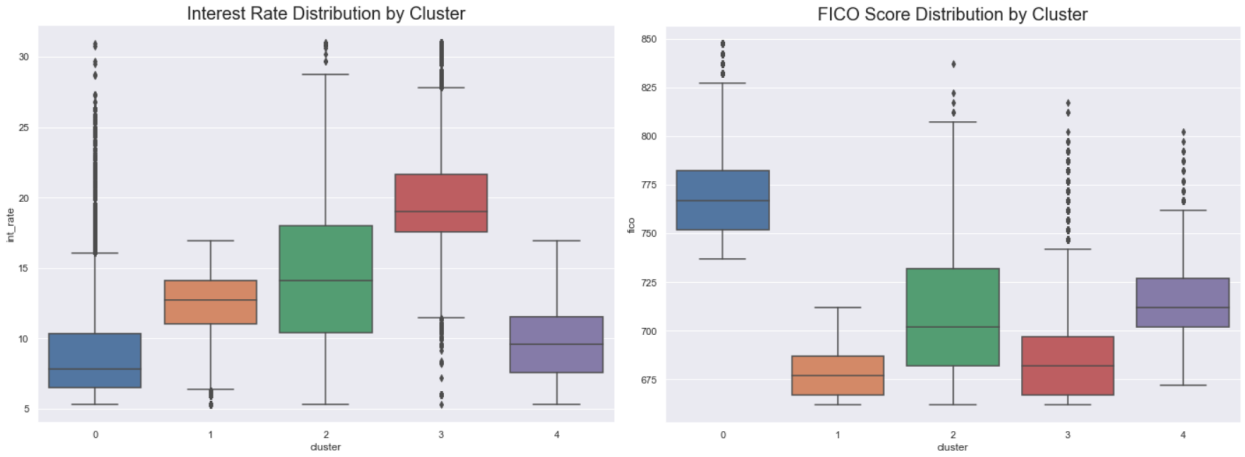


Figure 10: Cluster Centroids

Exploring these clusters further, we noticed major differences in their interest rate and FICO score distributions (Figure 11). Cluster 3 included loans with the highest overall interest rates and some of the lowest FICO scores. Meanwhile, cluster 0 had loans with the lowest overall interest rates and highest overall FICO scores. When examining whether the 5 clusters were distributed geographically differently, we found that the average rural proportion was about equal for all clusters, indicating that no particular cluster was more or less concentrated in rural areas.



(a) Interest Rate Distribution Across Clusters (b) FICO Score Distribution Across Clusters

Figure 11: Cluster Exploration

To analyze geographic discrimination within clusters, for each one, we calculated the correlation coefficient between the rural percentage of a loan's state population against that loan's interest rate (Figure 12).

Cluster Number	Correlation Coefficient
0	-0.041
1	0.189
2	-0.193
3	-0.048
4	0.535

Figure 12: By-Cluster Rural Proportion to Interest Rate Correlation Coefficients

While the correlation within clusters 0 and 3 was very weak, cluster 1 saw a weakly positive correlation, cluster 2 saw a weakly negative correlation, and cluster 4 saw a moderately strong positive correlation. Cluster 4 was characterized by high loan grades (predominantly A and B), mid-to-low FICO scores, and mid-to-low debt-to-income ratios. Our analysis indicates that within this group of loans, those that come from more-rural states tend to have higher interest rates.

By making comparisons between loans with similar grades, we ensure that the rural effect on interest rates is not confounded by the strong relationship between loan grades and interest rates. Thus, this result is particularly strong. After visualizing cluster 4 interest rates on a choropleth map, our results were further corroborated, as rural America generally observed higher interest rates than average (Figure 13).

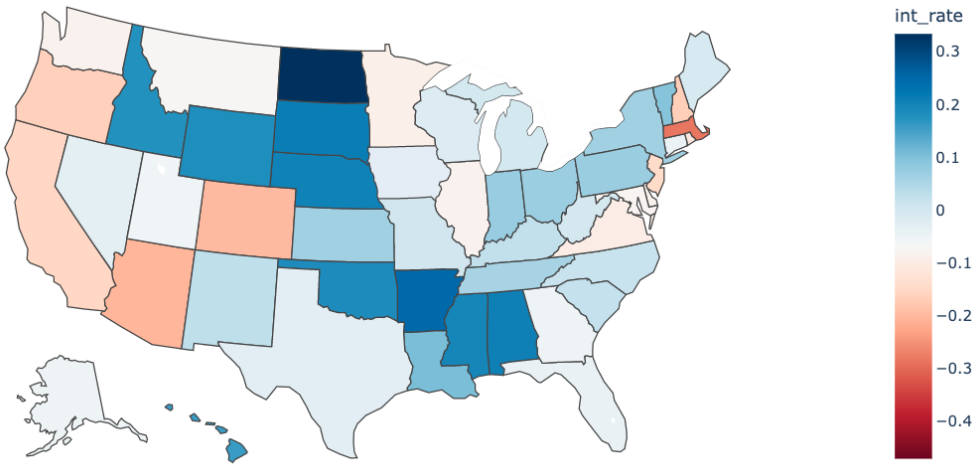


Figure 13: Cluster 4 - Interest Rates Across States Minus Average Interest Rate

To put these results into the framework of causal inference, we created a binary variable to classify if a loan is from a "more rural" area (i.e. if its rural proportion is above the median). We then estimated the average treatment effect $E[\tau_1]$ using the below formula.

$$\widehat{E[\tau_1]} = \hat{\theta}_1 - \hat{\theta}_0,\ \hat{\theta}_0 = \frac{\sum_{j=1}^{n} Y_j(1-w_j)}{\sum_{j=1}^{n}(1-w_j)},\ \hat{\theta}_1 = \frac{\sum_{j=1}^{n} Y_j w_j}{\sum_{j=1}^{n} w_j}$$

$\hat{\theta}_0$: sample mean for control (urban) group
$\hat{\theta}_1$: sample mean for treated (rural) group
Y_j: interest rate for sample j
w_j: binary indicator for if sample j comes from a rural population
n: total number of samples

We found that the estimated average effect on interest rates of a loan coming from a more-rural state was about 0.109 with a standard error of about 0.014. We could therefore infer that within this cluster of loans, a loan's state being more-rural had a moderately positive causal effect on the loan's interest rate.

3.5 Pre- and Post-2016 Analysis

Given the results of the clustering, we were curious whether the forced resignation of Renaud Laplanche, LendingClub's CEO, in 2016 had an effect on the discriminative treatment of rural

states' applications. Using the dataset curated in 3.1.1 Zipcode Demographics, we visualized the correlations, distributions, and scatterplots pairwise of each quantitative variable. To account for multicollinearity, we iterated through the visualizations, removing correlated and logically equivalent variables (% population male and female), applying log and logit transformations to normalize variables as needed. Results for the pre-2016 data is represented in Figure 14.

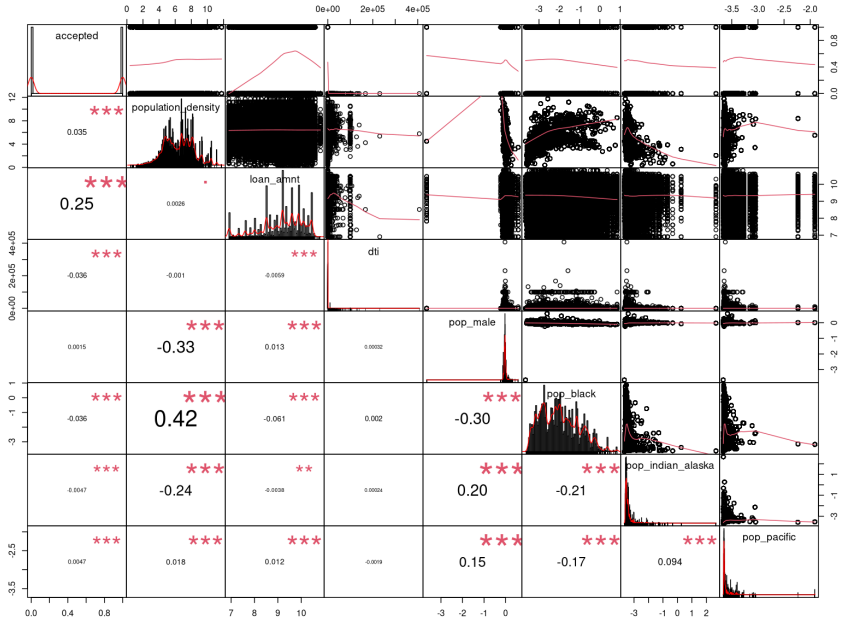


Figure 14: Example Final Pre-2016 Resignation Plot after Transformation

Logistic regression to predict acceptance of a loan application was performed twice for both pre-2016 resignation and post-2016 resignation: once with *Population Density* and once without. We analyzed the model prediction accuracy and found the difference in % accuracy, averaged across state, between the model with *Population Density* as a factor and the one without. Figures 15 and 16 depict the distribution of the impact of population density on model accuracy per state.

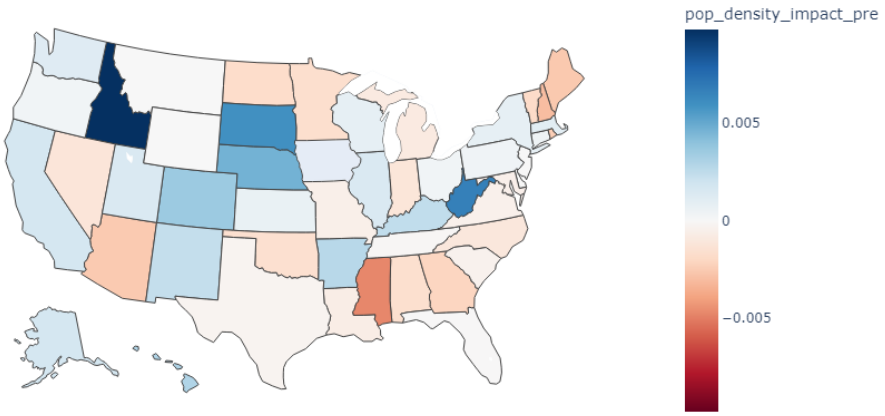


Figure 15: Impact of Population Density, Percentage Diff in Prediction Accuracy by State Post-2016

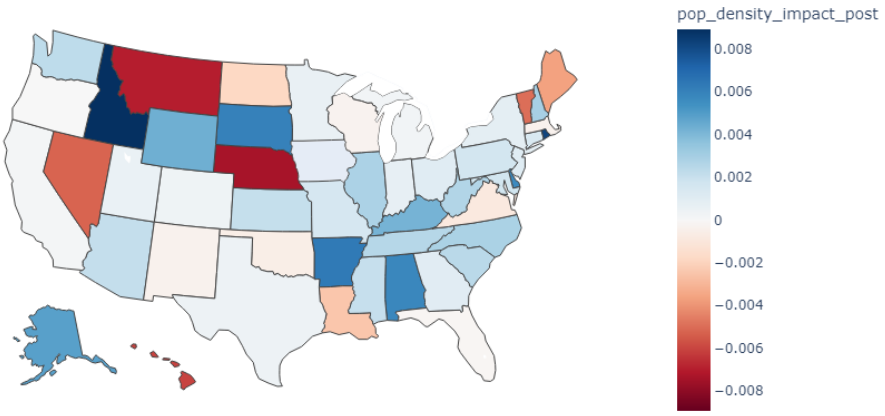


Figure 16: Impact of Population Density, Percentage Diff in Prediction Accuracy by State Post-2016

In particular, we found it interesting how population density seemed to have play a much milder role in predicting acceptance pre-2016 than post-2016. While the visualizations highlight these results, more robust analysis is needed. Other interesting notes we found:

1. The coefficient of log(*Population Density*) increased from 0.03 to 0.05 in our logistic regression, suggesting that increase in log(*Population Density*) made acceptance more

likely after the resignation in 2016.

2. Values of log(*Population Density*) had a median of about 6.49 for both pre- and post-resignation, indicating the scale of the variable was in the single digits, so a coefficient of 0.03 or 0.05 suggests a relatively small shift in application acceptance chance.

A possible insight into why LendingClub may have discriminated against rural applications comes from the patterns of peer-to-peer lending winding down post 2016. Evidenced by the trend of institutional investors taking over the funding market culminating in a 2020 closure of peer-to-peer investing to individuals, LendingClub has long sought to originate larger loan volumes [3]. More rural areas have less cash flow than urban economic hubs, and our geographical analysis demonstrated that rural states, particularly in the southeast, had higher proportion of loans with low grades than the rest of the country. Consequently, LendingClub could possibly have been more risk-averse when choosing loans to accept in order to satisfy institutional investors with safer returns, thereby discriminating against rural applications.

3.6 Conclusions

We found that for a particular cluster of loans characterized by high grade, mid-to-low FICO score, and mid-to-low debt-to-income ratio, the rural proportion of the loan's state correlated strongly with interest rates and had a moderately positive causal effect. In our analysis of how discrimination changed pre- vs. post-2016, we found that population density played a milder role in predicting loan acceptance pre-2016 than post-2016, although our results warrant further and more rigorous analysis.

References

[1] US Census Bureau. 113th Congressional District Summary File. `https://www.data.census.gov/`, Dec 2010.

[2] Max Chafkin and Noah Buhayar. How lending club's biggest fanboy uncovered shady loans. `https://www.bloomberg.com/news/features/2016-08-18/how-lending-club-s-biggest-fanboy-uncovered-shady-loans`, Aug 2016.

[3] Ruby Hinchliffe. Lendingclub shuts retail p2p offering as it focuses on institutional investors. `https://www.fintechfutures.com/2020/10/lendingclub-shuts-retail-p2p-offering-as-it-focuses-on-institutional-investors`, Oct 2020.

[4] Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. Jan 1990.

[5] Ben Luthi. Cfpb details financial challenges for rural communities. https://www.investopedia.com/cfpb-details-financial-challenges-for-rural-communities-5235826, Apr 2022.