Text, Demography, and Geography: An Analysis of Lending Club

Team 23 Submission - Summer Invitational 2022

Table of Contents

1 Topic Questions

1. How does social disclosure impact interest and default rates?
2. How are immigrants impacted by LendingClub and its 2016 changes?
3. How important is geography for LendingClub's growth?

2. Introduction

LendingClub, founded in 2006, is Peer-to-Peer lending service that allowed borrowers to create unsecured personal loans and investors to browse and select loan listings that fit their investment portfolio based on borrowers' and loan information. LendingClub grew extremely rapidly, quickly becoming the world's foremost Peer-to-Peer lender. In 2014, LendingClub became the biggest technology IPO of the year. In this report, we investigate the geographic growth of LendingClub in the United States using spatial regression.

However, in 2016 LendingClub became embroiled in controversy. The controversy included issues such as loan alterations, loans sold without meeting proper criteria, and possible discrimination on the basis of race. Following the controversy, LendingClub's popularity waned until 2020, when LendingClub closed its Peer-to-Peer lending platform. In this report, we sought to dive deeper into the effects of the controversy, namely, how LendingClub's model changed in 2016. Prior to the 2016 discrimination controversies, LendingClub had allowed borrowers to add descriptions to their loan applications to add further context, but removed this feature after 2016. We thus seek to investigate how social disclosure may help or harm a borrower's expected interest and default rates, and the manner in which this occurs.

The Peer-to-Peer lending platform was an attractive alternative to traditional sources of money lending, such as banks and credit cards, for many communities. Of these communities, we focused on immigrants. Immigrants often struggle to obtain opportunities to obtain money through traditional avenues because metrics such as credit score and length of employment are difficult to build over a short period of time. Thus, the 2016 removal of a description box and race features are likely to have an impact on immigrants, so we will investigate the relationship between immigrant population and the interest rate and prior to and after the 2016 changes.

2 Analysis on loan description provided by borrowers

Being accused of implicitly providing borrowers' demographic information through a description attached to their loan application, Lending Club no longer offered the opportunity for users to attach an explanation for their loan application, leading us to propose the question: what is the underlying semantic information of borrowers' loan description?

2.1 Exploratory data analysis on loan description

We started the analysis by investigating whether the presence of description has an influence on the default rate and interest rate of the loan. After grouping the data based on the loan status (fully paid and all other status), we created the table below to demonstrate the number of entries in each group, number of loans with descriptions and the proportion of loans with descriptions:

Paid	#Total	#Description	%Description
TRUE	715126	12823	0.017931
FALSE	713493	2517	0.003528

Table 2.1 Descriptions Summary, Grouped by Loan Status

While exploring the dataset, we noticed that in different LC assigned loan grade groups, the proportions of borrowers writing descriptions are slightly different. Thus we further grouped the dataset by grade groups:

Grade	Paid	#Toal	#Description	%Description
A	FALSE	125318	136	0.001085
A	TRUE	155694	2449	0.015730
B	FALSE	186225	538	0.002889
B	TRUE	219362	4541	0.020701
C	FALSE	213001	817	0.003836
C	TRUE	196322	3386	0.017247
D	FALSE	113727	628	0.005522
D	TRUE	90893	1649	0.018142
E	FALSE	53898	272	0.005047
E	TRUE	39068	587	0.015025
F	FALSE	16017	94	0.005869
F	TRUE	10816	171	0.015810
G	FALSE	5291	32	0.006048
G	TRUE	2971	40	0.013463

Table 2.2 Descriptions Summary, Grouped by Loan Status & Grade

2.1.1 Correlation between description and default rate

To investigate if there is a correlation between borrowers providing a description and fully paying the loan, we conducted a chi-square test. Since the p-value is close to 0, we concluded that there is statistically significant deviation in the default rate between the loans with and without descriptions.

We also grouped the dataset by LC assigned loan grade groups and conducted chi-square tests on each group. The results show that for all the grade groups, the statistical correlation presents, and this is especially true for grade groups A, B, and C. It is also noticeable that as Grade goes from A to G, p-value increases. In other words, for riskier loans, the correlation between description and default rate is not as significant as the safer groups. We believe that this finding could be attributed to the nature of riskier loans, which is more complex than that of safer loans, which means that whether a riskier loan is fully paid depends on more factors. The p-values of the chi-square tests are listed below:

Grade	P-Value
A	0.000000E+00
B	0.000000E+00
C	0.000000E+00
D	8.807464E-161
E	2.793605E-55
F	1.102595E-15
G	7.885460E-04

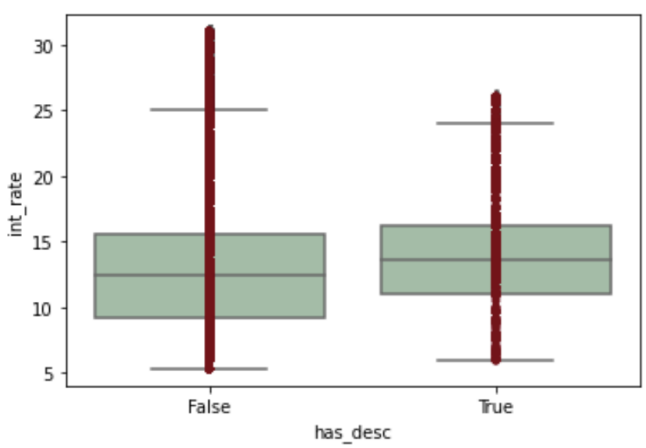
Table 2.3 Chi-square Test P-Values, Grouped by Grade

2.1.2 Correlation between description and interest rate

Because of the findings above, we speculated that there might be a correlation between description and loan's interest rate. We used the ANOVA test to test our hypothesis. We started off by calculating mean, median and standard deviation of the interest rates among loans with and without description. A box-plot was created to visualize the difference in distributions.

has_desc	Mean	Median	Std
FALSE	12.931666	12.49	4.788618
TRUE	13.721537	13.65	4.129342

Table 2.4 Interest Rate Summary



Graph 2.1 Interest Rate Distributions

Based on ANOVA test results, we concluded that there is statistical correlation between description and interest rate. The test result is reported as:

	sum_sq	df	F	PR(>F)
has_desc	9.467813E+03	1.0	414.024947	5.024551E-92
Residual	3.266887E+07	1428601.0	NaN	NaN

Table 2.5 ANOVA Test Result

Similar to what we have done for default rate, we also grouped the dataset by LC assigned grades to see if there are correlations between description and interest rate in different grade groups. Again, we first calculated the mean, median and standard deviation, and then used the ANOVA test to determine if a correlation exists for each group. The result suggests that for all groups, there is correlation between description and the interest rate and that correlation is the most significant for Grade A, B, and C. However, different from what we found earlier, p-values don't show any trends as grades go from A to G.

grade	has_desc	mean	median	std
A	FALSE	7.045037	7.21	0.957158
A	TRUE	7.892727	7.90	0.908570
B	FALSE	10.490502	10.49	1.196773
B	TRUE	11.801018	11.99	1.259627
C	FALSE	13.943879	13.98	1.207614
C	TRUE	14.837409	14.64	0.773135
D	FALSE	17.819420	17.57	1.659609
D	TRUE	18.047703	18.25	1.019464

E	FALSE	21.246903	20.49	2.605391
E	TRUE	21.412328	21.18	1.056126
F	FALSE	25.188196	24.50	2.605882
F	TRUE	24.285094	24.08	0.608007
G	FALSE	28.329491	28.14	2.193785
G	TRUE	25.879444	25.83	0.089504

Table 2.6 Interest Rate Summary, Grouped by Grades

Grade	P-Value
A	0.000000E+00
B	0.000000E+00
C	0.000000E+00
D	5.774249E-11
E	6.296213E-02
F	1.717552E-08
G	3.390101E-21

Table 2.7 ANOVA Test P-Values, Grouped by Grades

2.2 Description content

We were also interested in the influence that the description contents have on default rate and interest rate. Using Natural Language Processing techniques, we were able to extract a series of features from the descriptions which were later used for building regression models.

2.2.1 Data preparations and features

The first task was to clean up the description data. We followed the steps of (1) removing non-relevant information (i.e. the “Borrower added on …” prefix), (2) removing the punctuations, (3) transforming all words into lowercase, (4) removing stopwords[1] and (5) lemmatization. After each step, a new column was created because we anticipated that different features would need different levels of data cleaning.

Data Name	Level of Cleaning
desc	Original description data
desc_clean	Non-relevant information removed

[1] The stopword list we used is from nltk.corpus

desc_rem	Punctuations removed, transformed into lower-case
desc_sw_rm	Stopwords removed
desc_lem	Lemmatized

Table 2.8 Cleaned-up Data Info

Once the data was ready, we worked on building the features. The features we used and the data they were calculated from are listed below:

Feature Name	Feature Description	Data
#words	Number of words in the description	desc_clean
has_typo	Whether the description has typo	desc_rem
has_slang	Whether the description uses slangs and/or abbreviations	desc_rem
#grammar err	Number of grammar errors	desc_clean
#2nd person	Number of 2nd person pronouns used	desc_rem
#1st person plural	Number of plural 1st person pronouns used	desc_rem
pos	Number of adjectives and adverbs used	desc_rem
emotion	The strongest emotion conveyed	desc_rem
#family	Number of family related keywords	desc_lem
#debt	Number of debt related keywords	desc_lem
#business	Number of business related keywords	desc_lem

Table 2.9 Features Info

2.2.1.1 #words, #2nd person, #1st person plural

The “#words” feature is simply counting the total number of words in the original description. We suspected that longer descriptions are more advantageous over the shorter ones because they can convey more information. The “#2nd person” feature is the number of second person pronouns (i.e. lenders, you, your, etc.). Our hypothesis is that by referring directly to the lenders, the borrowers demonstrate their level of urgentness while at the same time show respect to the lenders. The “#1st person plural” feature is the number of plural first person pronouns (i.e. we, us, etc.), which gives information about if the beneficiary is limited to the borrower himself/herself and covers a wider range of people.

2.2.1.2 has_typo, has_slang, #grammar err

We believe that the writing habits, especially errors made in writings, demonstrate certain traits of the authors. For instance, having typos or grammar errors in the descriptions

indicates that the borrower is very likely to be careless or poorly-educated. The usage of slang and abbreviations is also a representation of the borrower's education level, and it also implies whether the borrower takes the loan seriously.

To calculate the "has_typo" feature, we first used the `spellchecker` library from python to count the number of words that may be misspelled. To calculate the "has_slang" feature, we scraped a list of acronyms and abbreviations and compared each word in the description with the list. The "#grammar err" feature was created with the help of the `language_tool_python` library.

2.2.1.3 pos

We learned from a NLP research that the number of adjectives and adverbs can be seen as indicators of subjectivity of the text (Rittman et al.). Taking this into account, we did part-of-speech analysis on the description data and counted the number of adjectives and adverbs as a way of evaluating its subjectivity. The part-of-speech analysis is conducted using the `nltk` library.

2.2.1.4 emotion

We also conducted emotion analysis on the descriptions. Considering the high level of similarities among the descriptions, compared to categorizing descriptions into two categories based on positive or negative sentiment, we believe that a more detailed emotion analysis would provide more useful insights about the description contents. We used a pre-trained emotion analysis model `EmoRoBERTa`[2], which is built on BERT's language masking strategy but trained on a larger dataset than BERT. Given a text, the model returns its strongest emotion. Applying the model on our description data gives the following categorization (we only selected the top 10 most common emotions here):

Emotion	Count
neutral	9029
approval	2972
gratitude	951
desire	883
caring	340
optimism	308
admiration	242
realization	172
sadness	100

[2] More information about the model can be found here

disappointment	59

Table 2.10 Emotion Analysis Result

2.2.1.5 #family, #debt, #business

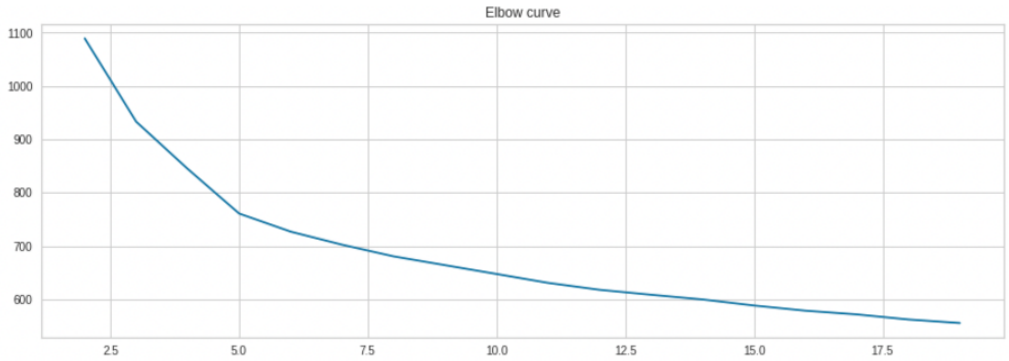
To dig deeper into the semantic information of the description data and derive further insights into the effect of different types of social disclosures on loan funding, we constructed a theme modeling machine learning pipeline that is comprised of four steps: sentence encoding, dimensionality reduction, text clustering, and a Counter to determine central words. We then counted the number of occurrences of the keywords in the descriptions and used them as features of the regression models.

The first step of the pipeline was to transform text into numeric vectors. There are multiple options to do this. For instance, we could adopt TF-IDF or Bag of Words to create representations based on word count. However, in these models the order of words is ignored and attention based models have proven to have higher performance in capturing the semantic meaning of texts. Hence, we decide to use the sentence-transformer `all-MiniLM-L6-v2`, which is trained specifically for semantic similarity tasks. Under the hypothesis, text vectors will position close to other similar texts and close to the most distinguishing words.

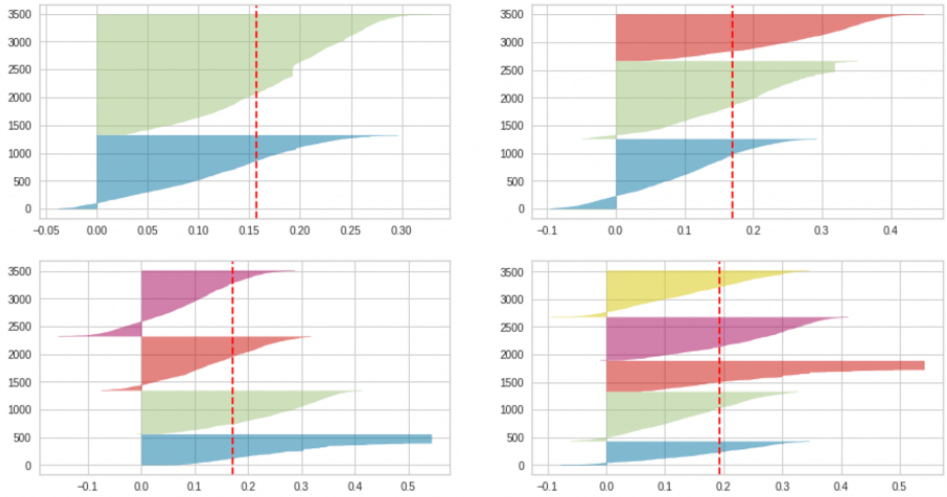
Next, given that our embedding vectors have a length of 384, we would need to map the high dimensional vector into a workable dimension space. Otherwise the distance metrics adopted by most clustering algorithms, such as the Euclidean distance, would become uniform for all vectors. Here we applied PCA, a light-weight algorithm that works well for short text with a small corpus if combined with k-means, which makes it perfectly aligning with our needs. We first chose the number of components to be 30, which explains up to 69.5% of the variance. Hyper parameter tuning was carried out in the following steps.

To determine the number of clusters to use in PCA, we first adopted the elbow method. Using cluster number from 2 to 20, it is clear to observe an elbow point around number 5. To obtain a more fine-grained result, we calculated the silhouette score for cluster number 3 to 8 to evaluate the quality of clusters created, which provides an indicator of how dense each cluster is and how well they are separate from each other. The higher the score, the more dense and well-separated each cluster is. From the graph we could arrive at the conclusion that 5 clusters is the optimum number for the k-means algorithm since it has the highest average score and also because the score for each cluster is above the average silhouette scores. In addition, the thickness of each cluster in the plot is approximately the same, proving that the fluctuation in size is similar.

To better fine tune the hyperparameters of the number of clusters and the number of principal components, we carried out a grid search (results in Appendix A). The final parameters are 10 for number of components and 5 for number of clusters.



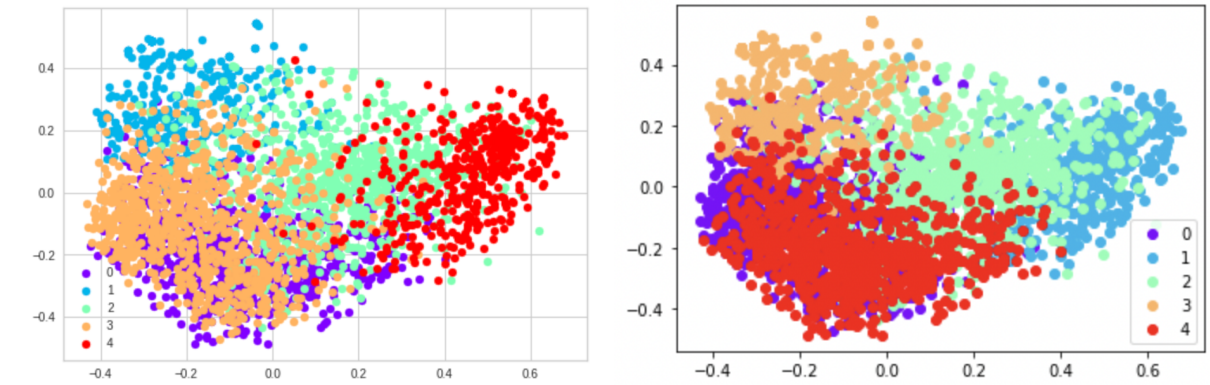
Graph 2.2 Elbow Curve



Graph 2.3 Silhouette Score on Cluster Sizes 2 to 5

Graph 2.4 Silhouette Score on Cluster Sizes 3 to 8

To visualize the clustering algorithm result, we further reduced the dimension of data points to 2d using PCA. The figure on the left represents descriptions of grade groups D to G and the figure on the right represents groups A to C.



Graph 2.5 Clustering Result of Grade D-G (left) & A-C (right) Description

Lastly we ran a class based variant of TF-IDF to obtain the main topics within each cluster. Instead of running TF-IDF on each sentence, which only obtains the information about which individual word has the most importance with respect to each description, this variant helps find the significance of each word to each cluster. We first joined all the descriptions in each class. Then, the frequency of each word was extracted for each class and divided by the total number of words. Next, the total unjoined number of documents across all classes was divided by the total frequency of words across all classes.

$$c - TF - IDF_i = \frac{t_i}{w_i} \times \log \frac{m}{\sum_j^n t_j}$$

Consistent with the plot above, each cluster contains a great deal of overlapping keywords such as *['consolidation', 'debt', 'credit', 'card', 'bills', 'loan', 'monthly', 'payoff']*.However, some distinctions were still spotted. In cluster 0, the most significant words contain *['home', 'improvement', 'medical', 'house', 'repair', 'kitchen', 'remodel', 'roof', 'car', 'vacation', 'bathroom', 'wedding', 'family']*. On the other hand, in cluster 4, the most frequent words include *['cycle', 'event', 'expense', 'budget', 'build' ,'business'].*

Given the findings above, we decided to look at three big categories – family (with keywords), debt (with keywords) and business (with keywords). By counting the number of times any keyword occurs in the description, we want to capture the purpose of the loans according to the borrowers and test whether it has a correlation with default rate and/or interest rate later on.

2.2.2 Correlation between description content and default rate

We then built a logistic regression model using all the features introduced above and looked at the p-values of each feature. Before modeling, we transformed the “emotion” feature and used numbers to replace the texts. The results are listed below:

	coef	std err	z	P>|z|	[0.025	0.975]
#words	0.0045	0.001	7.342	0.000	0.003	0.006

has_typo	0.0021	0.066	0.032	0.975	-0.127	0.131
has_slang	-0.1012	0.191	-0.530	0.596	-0.475	0.273
#grammar err	-0.0343	0.016	-2.151	0.032	-0.066	-0.003
#2nd person	-0.1164	0.062	-1.874	0.061	-0.238	0.005
#1st person plural	0.0556	0.045	1.230	0.219	-0.033	0.144
pos	-0.0193	0.016	-1.207	0.227	-0.051	0.012
#family	0.0124	0.028	0.438	0.661	-0.043	0.068
#debt	0.0949	0.014	6.639	0.000	0.067	0.123
#business	-0.1177	0.074	-1.595	0.111	-0.262	0.027
emotion_num	0.0617	0.002	29.914	0.000	0.058	0.066

Table 2.11 Logistic Regression Result

As the results show, “#words”, “#grammar err”, “#debt” and “emotion” features have statistically significant contributions to the model. However, surprisingly, the regression result shows that there is no correlation between the feature “has_typo” and the loan default rate. This suggests that spelling and grammar demonstrate perhaps different aspects of the borrowers’ background.

Remembering from previous findings that correlations between descriptions and default rate are more significant for Grade A, B and C, we ran three more logistic regressions. There are small differences among the results. For Grade A, the “#1st person plural” feature becomes significant, while “#grammar err” and “emotion” no longer have p-values less than 0.05.

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
#words	0.0119	0.003	4.514	0.000	0.007	0.017
has_typo	0.2365	0.276	0.857	0.391	-0.304	0.777
has_slang	-0.5440	0.660	-0.824	0.410	-1.837	0.749
#grammar err	-0.0173	0.074	-0.232	0.816	-0.163	0.128
#2nd person	-0.3774	0.223	-1.692	0.091	-0.815	0.060
#1st person plural	-0.3184	0.118	-2.700	0.007	-0.550	-0.087
pos	-0.0792	0.065	-1.226	0.220	-0.206	0.047
#family	-0.1481	0.109	-1.362	0.173	-0.361	0.065
#debt	0.0820	0.056	1.457	0.145	-0.028	0.192
#business	-0.3198	0.360	-0.888	0.375	-1.026	0.386
emotion_num	0.1144	0.008	13.608	0.000	0.098	0.131

Table 2.12 Logistic Regression Result, Grade A

For Grade B, “#words”, “#debt” and “emotion” are the only significant features.

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
#words	0.0055	0.001	4.188	0.000	0.003	0.008
has_typo	-0.1730	0.134	-1.294	0.196	-0.435	0.089
has_slang	0.4145	0.452	0.917	0.359	-0.471	1.300
#grammar err	-0.0136	0.035	-0.387	0.699	-0.082	0.055
#2nd person	-0.1742	0.125	-1.393	0.164	-0.419	0.071
#1st person plural	-0.0542	0.084	-0.642	0.521	-0.220	0.111

pos	0.0050	0.035	0.144	0.886	-0.063	0.073
#family	0.0511	0.065	0.785	0.432	-0.076	0.179
#debt	0.1336	0.030	4.389	0.000	0.074	0.193
#business	-0.1828	0.165	-1.108	0.268	-0.506	0.140
emotion_num	0.0790	0.004	18.005	0.000	0.070	0.088

Table 2.13 Logistic Regression Result, Grade B

For Grade C, “#words”, “#grammar err”, “#1st person plural” and “emotion” are significant.

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
#words	0.0043	0.001	3.970	0.000	0.002	0.006
has_typo	0.1860	0.122	1.525	0.127	-0.053	0.425
has_slang	0.0030	0.341	0.009	0.993	-0.666	0.672
#grammar err	-0.0921	0.031	-2.968	0.003	-0.153	-0.031
#2nd person	0.0152	0.124	0.123	0.902	-0.228	0.258
#1st person plural	0.2077	0.103	2.011	0.044	0.005	0.410
pos	-0.0456	0.028	-1.648	0.099	-0.100	0.009
#family	0.0762	0.051	1.495	0.135	-0.024	0.176
#debt	0.0895	0.025	3.544	0.000	0.040	0.139
#business	-0.0780	0.128	-0.611	0.541	-0.328	0.172
emotion_num	0.0548	0.004	14.931	0.000	0.048	0.062

Table 2.14 Logistic Regression Result, Grade C

2.2.3 Correlation between description content and interest rate

Finally, we used a linear regression model with interest rate as the dependent variable. Different from what we did with the logistic regression, we applied one-hot encoding on the “emotion” variable to look at how each individual emotion affected the model. The results are listed below (we only kept the emotions with low p-values here):

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	13.6865	0.281	48.640	0.000	13.135	14.238
#words	-0.0003	0.001	-0.369	0.712	-0.002	0.001
has_typo	-0.0786	0.100	-0.787	0.431	-0.274	0.117
has_slang	-0.1203	0.299	-0.403	0.687	-0.706	0.465
#grammar err	0.0604	0.023	2.662	0.008	0.016	0.105
#2nd person	-0.1381	0.121	-1.145	0.252	-0.374	0.098
#1st person plural	-0.0412	0.059	-0.698	0.485	-0.157	0.075
pos	-0.0729	0.023	-3.205	0.001	-0.117	-0.028
#family	0.2738	0.042	6.596	0.000	0.192	0.355
#debt	-0.1339	0.021	-6.443	0.000	-0.175	-0.093
#business	0.7459	0.112	6.660	0.000	0.526	0.965
caring	0.9164	0.350	2.621	0.009	0.231	1.602
neutral	0.5031	0.278	1.811	0.070	-0.041	1.048

sadness	0.8652	0.490	1.767	0.077	-0.095	1.825

Table 2.15 Linear Regression Result

For interest rate, we also repeated the regression on Grade A, B and C. We noticed that in different groups, different emotions become significant. For Grade A, the emotion “anger” is statistically significant.

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	13.6865	0.281	48.640	0.000	13.135	14.238
#words	-0.0003	0.001	-0.369	0.712	-0.002	0.001
has_typo	-0.0786	0.100	-0.787	0.431	-0.274	0.117
has_slang	-0.1203	0.299	-0.403	0.687	-0.706	0.465
#grammar err	0.0604	0.023	2.662	0.008	0.016	0.105
#2nd person	-0.1381	0.121	-1.145	0.252	-0.374	0.098
#1st person plural	-0.0412	0.059	-0.698	0.485	-0.157	0.075
pos	-0.0729	0.023	-3.205	0.001	-0.117	-0.028
#family	0.2738	0.042	6.596	0.000	0.192	0.355
#debt	-0.1339	0.021	-6.443	0.000	-0.175	-0.093
#business	0.7459	0.112	6.660	0.000	0.526	0.965
anger	0.9164	0.350	2.621	0.009	0.231	1.602
grief	0.5031	0.278	1.811	0.070	-0.041	1.048
pride	0.8652	0.490	1.767	0.077	-0.095	1.825

Table 2.16 Linear Regression Result, Grade A

For Grade B, the emotions “curiosity” and “relief” are significant.

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	11.7580	0.156	75.232	0.000	11.452	12.064
#words	-0.0007	0.000	-1.460	0.144	-0.002	0.000
has_typo	-0.0467	0.053	-0.884	0.377	-0.150	0.057
has_slang	0.0518	0.164	0.316	0.752	-0.269	0.373
#grammar err	0.0091	0.011	0.814	0.416	-0.013	0.031
#2nd person	0.0214	0.063	0.340	0.734	-0.102	0.145
#1st person plural	0.0044	0.031	0.144	0.885	-0.055	0.064
pos	0.0069	0.012	0.568	0.570	-0.017	0.031
#family	0.0482	0.024	2.013	0.044	0.001	0.095
#debt	0.0033	0.011	0.296	0.767	-0.019	0.025
#business	0.0542	0.061	0.881	0.378	-0.066	0.175
curiosity	-2.0399	0.744	-2.740	0.006	-3.499	-0.580
grief	-5.83E-16	3.19E-16	-1.829	0.068	-1.21E-15	4.2E-17
relief	-3.007E-16	1.44E-16	-2.086	0.037	-5.83E-16	-1.81E-17

Table 2.17 Linear Regression Result, Grade B

For Grade C, the emotion “annoyance” is significant.

	coef	std err	t	P>|t|	[0.025	0.975]
const	14.7351	0.114	129.265	0.000	14.512	14.959
#words	-3.161E-05	0.000	-0.095	0.924	-0.001	0.001
has_typo	-0.0197	0.037	-0.535	0.593	-0.092	0.052
has_slang	0.0472	0.104	0.455	0.649	-0.156	0.251
#grammar err	0.0023	0.009	0.239	0.811	-0.016	0.021
#2nd person	-0.1077	0.049	-2.200	0.028	-0.204	-0.012
#1st person plural	0.0204	0.023	0.889	0.374	-0.025	0.065
pos	0.0023	0.008	0.272	0.786	-0.014	0.019
#family	-0.0134	0.015	-0.913	0.361	-0.042	0.015
#debt	-0.0094	0.008	-1.247	0.212	-0.024	0.005
#business	0.0291	0.040	0.731	0.465	-0.049	0.107
annoyance	-0.5931	0.281	-2.113	0.035	-1.143	-0.043
caring	0.2382	0.138	1.728	0.084	-0.032	0.508

Table 2.18 Linear Regression Result, Grade C

2.3 Conclusion

As a way of directly communicating to the lenders, descriptions written by the borrowers provide valuable information about borrowers themselves as well as their needs. In Section 2, we focused on analyzing the descriptions provided by borrowers. By setting the dependent variable as default rate and interest rate, we were able to conduct hypothesis testing and regressions. While we tried our best to come up with features that might have an influence on the dependent variables, the result turned out to be surprising. If we had the opportunity to analyze the data further, we'd like to investigate (1) how typos and grammar errors differ in terms of demonstrating author's backgrounds and personality traits and (2) how different emotions expressed in the descriptions are associated with default rate and/or interest rate. These require understanding of Natural Language Processing and linguistics concepts in depth, as well as adequate supplement datasets to train relevant models.

3. Are immigrants at a disadvantage?

Immigrants, comprising over 44 million people in the United States in 2018, are a growing population within the United States. Since immigrants have spent less time in the United States, it can be difficult to build credit history and generate numerical financial statistics that represent their status accurately. However, immigrants are a critical user group in Peer-to-Peer lending. Because immigrants are likely to be disadvantaged by quantitative factors which are more heavily used by traditional lending avenues, Peer-to-Peer lending offers a unique opportunity to get loans that would not be approved through traditional means.

In 2016, LendingClub removed data about a borrower's race due to concerns about racial discrimination. This form of discrimination may manifest in many ways in Peer-to-Peer lending, for example, a lender may choose to accept more loans from a certain racial group, or choose to impose a higher interest rate for a borrower of a certain race. LendingClub also removed the ability for borrowers to leave a description message, so a possible way for a borrower to explain extenuating circumstances was removed. Immigrant communities are

largely composed of underrepresented and marginalized groups, removing race from the available data might benefit immigrants by allowing them to avoid race-based prejudice. However, it is also possible that there is very little or even a backwards effect of the 2016 change, as lenders who see an in-group connection with others of a specific race and background are more likely to take on those loans, even if they are less well aligned with their investment goals (Geven). By removing the descriptions, immigrants would lose a possible avenue to explain their situation, for example recent immigrants are likely to have a low credit score due to having new accounts. Instead, immigrants are left with the same issue as if they were to apply for loans from traditional avenues, where they are forced to rely solely on financial details.

Due to the importance of Peer-to-Peer lending within immigrant communities, we investigated the question: what is the effect of the percent of immigrants in a city on the mean interest rate of the city?

3.1 Methods

The Random Effects model is a hierarchical linear model that allows us to control for constant unobserved heterogeneity. Given that there is plenty of information not provided by the Lending Club, we would like to control dependencies of unobserved, independent variables on our dependent variable, or else we might lead to biased estimators if using traditional linear regression models. In order to use the Random Effects model, we verified the random effects assumption, which is that the individual unobserved heterogeneity is uncorrelated with the independent variables. We can use this model by regressing the immigrant population percentage and financial variables on interest rate to determine the relationship between the immigrant population percentage and interest rate.

To prepare our data for the Random Effects model, we acquired the percentage of the total population of a city that are immigrants, and matched the city to its corresponding zip code(s) to the accuracy of the first 3 numbers of the zip code. We used the city as our aggregation level because it was the finest level of granularity that was allowed for by the zip codes, yet would still be detailed enough to reflect the effects of immigrant presence. Then, for the metrics we chose to investigate, we were able to use the Lending Club's accepted loans dataset to match on the zip codes of the immigrant percentage dataset to generate a dataset with all of the financial information along with the total percentage of immigrants per city. In order to investigate the effect of the 2016 changes, the dataset was split between pre-2016 entries and post-2016 entries.

In our first analysis, we sought to investigate the relationship between the percentage of immigrants in a city and the interest rate of their application. We formulated the following null hypothesis: there is no relationship between the percentage of immigrants in a city and the mean interest rate of the city.

To investigate this hypothesis, we performed a Random Effects regression, where our dependent variable was the interest rate of each loan application,and the independent variable $Imm\,\%$ is the percentage of the number of immigrants in the city. The β_1 coefficients measures the effect of the magnitude of immigrant population percentage on the interest rate. To control for financial information the investors could obtain from the Lending Club, we use $FinanVar$,

which includes the FICO score, employment length, log of annual income, debt-to-income ratio, and the grade of loan, as we believed that it was a good summary metric of the financial data given. ϕ_t controls the yearly fixed effects and φ_i controls for the Metropolitan Statistical Area fixed effects. $\gamma_{i,t}$ controls for the combined yearly × Metropolitan Statistical Area fixed effect.

$$IntRate_{i,t} = \beta_0 + \beta_1 \times Imm\,\% + \beta_2 \times FinanVar + \phi_t + \varphi_i + \gamma_{i,t} + \varepsilon_{i,t}$$

3.2 Results

	R^2	P-value	β_1
Pre-2016	0.9249	0.0021	0.0442
Post-2016	0.9494	0.0249	0.0325

Table 3.1 Regression statistics produced by Random Effects model, with FICO score

In both the pre-2016 and post-2016 investigation, at the 5% significance level, we reject the null hypothesis since our p-value of 0.0249 does not exceed our significance level. The relationship between percent of immigrants in the city and mean interest rate in the city cannot be explained by random chance alone.

RandomEffects Estimation Summary

Dep. Variable:	int_rate	R-squared:	0.9249
Estimator:	RandomEffects	R-squared (Between):	0.9038
No. Observations:	962786	R-squared (Within):	0.9246
Date:	Sun, Jul 24 2022	R-squared (Overall):	0.9249
Time:	19:38:31	Log-likelihood	-1.588e+06
Cov. Estimator:	Unadjusted		
		F-statistic:	1.078e+06
Entities:	479	P-value	0.0000
Avg Obs:	2010.0	Distribution:	F(11,962774)
Min Obs:	1.0000		
Max Obs:	2.324e+04	F-statistic (robust):	1.078e+06
		P-value	0.0000
Time periods:	3	Distribution:	F(11,962774)
Avg Obs:	3.209e+05		
Min Obs:	2.084e+05		
Max Obs:	3.82e+05		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	10.675	0.0441	241.99	0.0000	10.588	10.761
Imm% tot pop	0.0442	0.0144	3.0800	0.0021	0.0161	0.0724
fico	-0.0043	4.748e-05	-89.790	0.0000	-0.0044	-0.0042
emp	0.0002	0.0004	0.4564	0.6481	-0.0005	0.0009
dti	0.0031	0.0002	20.359	0.0000	0.0028	0.0034
annual_inc	-0.1246	0.0060	-20.750	0.0000	-0.1364	-0.1129
grade.B	3.1978	0.0042	760.85	0.0000	3.1896	3.2061
grade.C	6.4865	0.0043	1494.0	0.0000	6.4780	6.4950
grade.D	10.147	0.0050	2025.2	0.0000	10.137	10.156
grade.E	13.305	0.0060	2200.3	0.0000	13.293	13.316
grade.F	17.341	0.0092	1892.9	0.0000	17.323	17.359
grade.G	20.118	0.0170	1181.5	0.0000	20.085	20.152

Table 3.2 Regression statistics produced by Random Effects model, pre-2016 with expanded financial variables

RandomEffects Estimation Summary

Dep. Variable:	int_rate	R-squared:	0.9494
Estimator:	RandomEffects	R-squared (Between):	0.9385
No. Observations:	813224	R-squared (Within):	0.9493
Date:	Sun, Jul 24 2022	R-squared (Overall):	0.9494
Time:	19:42:43	Log-likelihood	-1.277e+06
Cov. Estimator:	Unadjusted		
		F-statistic:	1.388e+06
Entities:	470	P-value	0.0000
Avg Obs:	1730.3	Distribution:	F(11,813212)
Min Obs:	1.0000		
Max Obs:	2.044e+04	F-statistic (robust):	1.388e+06
		P-value	0.0000
Time periods:	2	Distribution:	F(11,813212)
Avg Obs:	4.066e+05		
Min Obs:	3.848e+05		
Max Obs:	4.285e+05		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	11.111	0.0421	263.81	0.0000	11.028	11.193
Imm% tot pop	0.0325	0.0145	2.2435	0.0249	0.0041	0.0608
fico	-0.0045	4.265e-05	-105.06	0.0000	-0.0046	-0.0044
emp	-0.0032	0.0004	-8.9265	0.0000	-0.0039	-0.0025
dti	0.0031	8.288e-05	36.843	0.0000	0.0029	0.0032
annual_inc	-0.1675	0.0057	-29.455	0.0000	-0.1786	-0.1563
grade.B	3.5694	0.0038	937.84	0.0000	3.5619	3.5768
grade.C	7.2137	0.0040	1789.6	0.0000	7.2058	7.2216
grade.D	11.877	0.0049	2415.9	0.0000	11.868	11.887
grade.E	17.649	0.0072	2444.4	0.0000	17.635	17.663
grade.F	22.385	0.0133	1678.1	0.0000	22.359	22.411
grade.G	23.551	0.0196	1202.8	0.0000	23.513	23.590

Table 3.3 Regression statistics produced by Random Effects model, post-2016 with expanded financial variables

3.3 Conclusion

From our testing, we can see that there appears to be a relationship between immigrants and interest rate in a city after 2016 but not prior to 2016. Both parameters are positive, indicating that the percentage of immigrants has a positive correlation with the interest rate, and thus revealing the underlying bias. However, the parameter decreases after 2016. This could indicate that the factors –credit score, employment length, and so on–are more heavily observed in the post-2016 scenario. This makes sense with actual events, since after removing both race information and descriptions, the LendingClub landscape had to rely more on the financial information provided.

4 How important is geography for LendingClub's growth?

Next, we sought to explore what the significance of geographic distance was for the growth of LendingClub over time. As an online financial service company, it would be immediately sunclear whether LendingClub's growth comes from effective online marketing and targeting consumer traits, or whether the loan's location plays a part in influencing whether future loans will happen near it as well.

We used an expanded LendingClub approved loans dataset (from 2007 to 2018) to get a fuller picture of LendingClub's full range of growth over its lifetime. We also merged data from the US census, and Geocode (to get counties from trimmed ZIP codes and the distances between all pairs of counties).

4.1 Methods

Inspired by (Havrylychk, 2016), we decided to use a Spatial Autoregressive Model with Autoregressive Disturbances. In essence, the model regresses the loan volume for a county's month over its past month, the distance-weighted average of loan volume in other counties, and then controls for a variety of loan-specific and county-level factors.

The model is described below:

$$\mathrm{Y}_{i,t} = \alpha + \gamma\, \mathrm{Y}_{i,t\text{-}1} + \beta\, \mathbf{D}_i\, \mathrm{Y}_{\sim i,t} + \zeta\, \mathrm{W}_{i,t} + \delta\, \mathrm{X}_i + \mathrm{u}_{i,t}$$

Y (for a given county *i* and month *t*) is the log number of loans made in that county. α serves as the constant, γ is the coefficient for the one-month lagged log number of loans. D is a matrix of inverse distances between all pairs of cities; hence $D_i\, Y_{\sim i,t}$ for some given *i,t* is the inverse distance weighted average of loan volume in all counties besides county *i*. β is the coefficient for this, and we expect a positive sign on it if LendingClub loan activity grows geographically from one county to nearby ones. X is a matrix of all control variables on the county-level (not adjusted for each month), scraped from the US census. Such variables include demographic and socioeconomic variables found in the next section's table (all variables "TotalPop" onward); δ are the coefficients for these variables. W is a matrix of control variables that are time dependent (and thus from the LC data). They include FICO scores, hardship ratios, debt-to-income ratios, and average annual incomes of the borrowers who took the loans. u is the error term.

The null hypothesis is β = 0, and LendingClub's growth is not explainable by geographic proximity of loans over time.

4.2 Results

Dep. Variable:	log_loan_count	R-squared (uncentered):	0.967
Model:	OLS	Adj. R-squared (uncentered):	0.967
Method:	Least Squares	F-statistic:	8.466e+04
Date:	Sun, 24 Jul 2022	Prob (F-statistic):	0.00
Time:	13:14:48	Log-Likelihood:	-38627.
No. Observations:	48401	AIC:	7.729e+04
Df Residuals:	48384	BIC:	7.744e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
hardship_flag	-1.2818	0.623	-2.056	0.040	-2.504	-0.060
log_annual_inc	0.0488	0.008	5.752	0.000	0.032	0.065
dist_weighted_avg	0.2102	0.004	57.513	0.000	0.203	0.217
prev_log_loan_count	0.7937	0.002	335.748	0.000	0.789	0.798
fico_range_low	-0.0009	0.000	-6.657	0.000	-0.001	-0.001
dti	-0.0021	0.000	-4.531	0.000	-0.003	-0.001
TotalPop	6.63e-09	1.8e-09	3.675	0.000	3.09e-09	1.02e-08
Hispanic	0.0017	0.000	5.094	0.000	0.001	0.002
Black	0.0044	0.000	16.269	0.000	0.004	0.005
Native	-0.0027	0.001	-5.146	0.000	-0.004	-0.002
Asian	0.0040	0.001	6.770	0.000	0.003	0.005
Citizen	-0.5291	0.079	-6.662	0.000	-0.685	-0.373
Poverty	-0.0041	0.001	-5.409	0.000	-0.006	-0.003
Income	0.0541	0.011	4.711	0.000	0.032	0.077
Unemployment	0.0064	0.001	5.032	0.000	0.004	0.009
SelfEmployed	-0.0227	0.001	-15.677	0.000	-0.026	-0.020
FamilyWork	-0.1048	0.016	-6.379	0.000	-0.137	-0.073

Omnibus:	1871.366	Durbin-Watson:	1.672
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4482.747
Skew:	-0.218	Prob(JB):	0.00
Kurtosis:	4.426	Cond. No.	4.89e+08

Table 4.1 Regression Results

The row highlighted in blue shows a positive coefficient (statistically significant at the 0.0001 level) on the inverse-distance-weighted average of loans in all other counties, even with having the previous month's loan volume and a variety of demographic controls as other features in the regression. The model itself has a R-squared of 0.967, though shows a high degree of collinearity (as evidenced by the high condition number), pointing to unstable estimates.

4.3 Conclusion

From these results, we see the LendingClub loan activity for countries over time has a strong positive association with loan activity across all other counties. This points to (geographic) network effects that may not be unexpected in a peer-to-peer platform like LendingClub. People tend to find out about loans those around them are making and then participate in those loans themselves over time. As a company, LendingClub could have more aggressively marketed itself in strongly-leveraged geographies to expand into newer ones, in conjunction with other marketing strategies to grow lending activity on its peer-to-peer platform.

6 Conclusion

The study aims to investigate how the implicit non-financial information provided by Lending club influences the interaction between borrowers and investors and finally presents an overview of how geography positively shapes an online P2P platform like LendingClub's growth.

We answered three questions sequentially: First, we discovered correlation between loan description and interest rate as well as default rate. Then we dived deeper to identify a set of features of borrowers' loan description that were statistically significant to have effect on borrowers' interest rate and default rate, respectively, and it is advantageous for borrowers to invest time into well-written descriptions. Secondly, we specifically focus on immigrant borrowers, and arrive at the conclusion that there exist bias towards borrowers living in cities with high percentage of immigrants in terms of interest rate, and the series of acts of lending club after the 2016 scandal, including removing the zip code information and loan descriptions, has reduced the bias to some extent, given that the only metrics for measuring a client are financial indicators. Last, we turned our focus to Lending Club itself and analyzed its business model by inferring geolocation data. We came to the conclusion that geography does have a positive effect on the growth of lending, pointing at strong network effects presence even for online platforms.

These results have great implications on Peer-to-Peer lending and business growth. Our NLP analysis on borrowers' description may not have an impact on Lending Club algorithms, since the new regulation carried out in 2016 prohibits descriptions. However, other Peer-to-Peer lending platforms, such as Kiya, still depend heavily on borrower's descriptions as an informal way of allowing the borrowers to win investments and lenders to estimate borrowers' credibility. In that sense our analysis would be useful in providing insights about evaluating borrowers based on description. In the Peer-to-Peer lending scene, our results that immigrants may be disadvantaged due to the removal of descriptions and race metrics, which makes Peer-to-Peer lending more similar to traditional lending, indicates that while the efforts were made to reduce discrimination based on race or other similar factors, they may backfire by causing marginalized communities such as immigrants to struggle in this context. The evidence of geography as a significant factor in driving county-level growth shows that for peer-to-peer networks, even online ones like LendingClub, geography still plays a major positive role in expanding the use of a platform. As a future exploration, it could be interesting to investigate what degree of geographic growth purely-online platforms have.

7 Appendix

Appendix A

See code here:
https://drive.google.com/drive/folders/1uDjDcdP2FBNJYWdkSdXwu-rExJttVVsr?usp=sharing

Appendix B

Grid search for best hyperparameter combination

#*	#**	Score	#*	#*	Score	#*	#**	Score
3	10	0.021535900197972524	4	200	0.09971242943754605	6	100	0.12511141539146511
3	20	0.16999668531409493	5	10	0.25020103622953627	6	200	0.1135991180216649
3	30	0.15183939713448516	5	20	0.19314794911061295	7	10	0.23746083295519613
3	40	0.1417319480933951	5	30	0.17101382542329777	7	20	0.18565216362104664
3	50	0.1348071746820673	5	40	0.15828795506348303	7	30	0.16491882703062227
3	100	0.11848433300272045	5	50	0.14973276803652638	7	40	0.15331798974975316
3	200	0.10894222786033883	5	100	0.12941891774346206	7	50	0.14540636043063324
4	10	0.22282461109589732	5	200	0.11694170681793441	7	100	0.1265106056242639
4	20	0.17110468423387734	6	10	0.24053451466964976	7	200	0.11487254413853575
4	30	0.15041860105548677	6	20	0.1867165787690253	8	10	0.21593108712599934

4	40	0.13850487200512854	6	30	0.16516754440121112	8	20	0.18514804921437128
4	50	0.13055712336740516	6	40	0.15288033941320783	8	30	0.154365101720085
4	100	0.11151287661668391	6	50	0.1445505507387778	8	50	Score 0.1357690220104131

8 Citations

Chung, Matthew D. “Small Business Loans and Natural Language Processing.” *ProQuest Dissertations and Theses*, California State University, Los Angeles, 2020. *ProQuest*, https://www.proquest.com/docview/2557817969/abstract/BC74C157FF1B470DPQ/1.

Geven, Stef. *Taste-Based Discrimination: In-Group Bias on Peer-to-Peer Lending Platforms? Evidence from Lending Club*. p. 42.

Havrylchyk, Olena, et al. *What Drives the Expansion of the Peer-to-Peer Lending?* 2016.

Lee, Michelle Seng Ah, and Jatinder Singh. *Spelling Errors and Non-Standard Language in Peer-to-Peer Loan Applications and the Borrower’s Probability of Default*. 25 May 2020. *Social Science Research Network*, https://doi.org/10.2139/ssrn.3609834.

Reddy, Sriharsha, and Krishna Gopalaraman. “PEER TO PEER LENDING, DEFAULT PREDICTION-EVIDENCE FROM LENDING CLUB.” *The Journal of Internet Banking and Commerce*, vol. 21, no. 3, Nov. 2016. *www.icommercecentral.com*, https://www.icommercecentral.com/peer-reviewed/peer-to-peer-lending-default-predictionevidence-from-lending-club-81766.html.

Rittman, Robert, et al. “Adjectives as Indicators of Subjectivity in Documents.” *Proceedings of the American Society for Information Science and Technology*, vol. 41, no. 1, 2004, pp. 349–59. *Wiley Online Library*, https://doi.org/10.1002/meet.1450410141.

Shen, Yuanyuan. “Engagement Drivers in a Lending Marketplace: The Case for Kiva.” *ProQuest Dissertations and Theses*, Stanford University, 2019. *ProQuest*, https://www.proquest.com/docview/2466288975/abstract/14EC0053D511436DPQ/1.