UNIVERSITÉ LIBRE DE BRUXELLES, UNIVERSITÉ D'EUROPE

Solvay Brussels School
Economics & Management

MEMOIRE

Présenté en vue de l'obtention du Master en ingénieur de gestion

Big Data for Credit Scoring: towards the End of Discrimination on the Credit Market? Evidence from Lending Club

Par Orphée Van Essche

Directeur: Professeur Marek Hudon
Assesseur: Professeur Nicolas Van Zeebroeck

Année académique 2017 - 2018

Table of Contents

Abstract

Discrimination, especially towards racial minorities, is still pervasive on the credit market. Whether the use of big data for credit scoring by FinTech lending companies will limit or facilitate such discrimination remains an open debate in the literature. This research paper discusses this issue by searching for geographic exclusion of areas with a predominant African-American community in a new kind of credit market, peer-to-peer lending. Using data collected from the biggest US lending platform, Lending Club, matched with demographics from the US Census Bureau, we find evidence of taste-based discrimination. Borrowers from areas mainly populated in African-Americans have 20% lower probability to be accepted for a loan and pay a 20 basis points higher interest rate, whereas these disparities are not justified by a lower loan performance. These findings suggest that big data for credit scoring is not the optimal tool to erase discrimination from the credit market.

1 Introduction

With the rise of information and telecommunication technology, people are becoming more and more interconnected, exchanging always more data all around the globe. This increasing volume of information, transferred ever more rapidly under a variety of format, leaves behind a huge amount of digital trails. This is what we call *big data*, and which represents the new fuel of companies. Indeed, the same way old manufacturing companies depend on oil to ensure their production, more recent companies are now depending on data collected from their customers to enhance their service and product offering in order to survive in this increasingly competitive global market. This use of big data by companies to know and serve their customers better takes place in a variety of different field, such as in marketing where potential clients are now targeted with promotions more accurately tailored to their needs. However, big data is also used in an area that is highly determinant in people life opportunities: the credit market. Credit scoring, which is usually the first stage in loans' approval and pricing, and which depends on credit risk factors, may nowadays be complemented by non-traditional data about loans applicants such as browser data, social network and mobile data.

Big data for credit scoring enables FinTech loan companies to increase the predictive power of their scoring algorithms, which allows them to save on costs since, by predicting better which applicants will engage in risky behaviors, these big data companies can significantly decrease the default rate of borrowers. Moreover, by adding non-traditional credit risk factors about loan applicants to their scoring algorithms, these FinTech lenders can now extend their credit offer to traditionally underserved communities with few or no reported credit history, but who use a lot their smartphone and the internet. Beyond the considerable opportunities big data brings in term of efficiency in the scoring procedure, the debate related to the fairness it brings on the credit market is much more nuanced. More automated procedures incorporating big data into scoring algorithms could limit human discretion in the loan approval and pricing process. Such use of big data for credit scoring could then help prevent lenders from using prohibited characteristics such as race or gender in the scoring procedure, so that applicants would be assessed based on more neutral factors instead. However, besides its potential for lower reliance on human bias on the credit market, big data for credit scoring could also be used by lenders to exclude protected groups of the population from accessing credit in a totally concealed manner. Indeed, since big data is used to better profile individuals, lending companies could also make use of it to create proxies of loan applicants'

attributes that are prevented from using by law, such as ethnicity. For instance, the way an individual uses social media can be an indicator of his ethnicity. Then, such proxies could be included in the algorithms of these FinTech lenders so that the outcome of the scoring procedure would deny access to credit more to minority groups, such as African-Americans. As a result, this discrimination would occur without any prohibited attributes being directly incorporated in the scoring algorithm. Moreover, since these big scoring algorithms are proprietary, such fraudulent procedure would be hard to uncover.

To help fill this research gap regarding whether the use of big data for credit scoring would either alleviate or increase exclusion from the credit market of already discriminated minorities, we narrow our study on the online peer-to-peer lending market. This new credit marketplace, where investors and borrowers are brought together through an intermediate platform, is particularly appropriate for such analysis since these platforms use non-traditional data about loan applicants to determine what loans to accept and on what terms. In particular, we search for the presence of discrimination towards geographic areas mainly composed of African-Americans in the biggest US online peer-to-peer platform, Lending Club, whose part of data about loan applicants are made publicly available. In these public data, one can find the first three digits of a loan applicant's zip code, which we use, alongside with the demographic data from the US Census Bureau, to find for each borrower the proportion of African-Americans that are present in his housing area. With this information, we then analyze to what extent the specific ethnic density of a borrower's living area impacts the probability of that borrower to be accepted for a loan as well as the interest rate he is charged on his loan.

Our results show that, after controlling for observable credit characteristics, borrowers that live in areas where the African-American community exceeds 50% of the population have 20% less probability to be accepted for a loan compared to the other borrowers. Moreover, once accepted on the online platform, these same borrowers are charged an interest rate on their loan that is 20 basis points higher than for the other group of borrowers. Our findings indicate also that this discrimination towards borrowers living in areas densely populated in African-Americans on the online peer-to-peer platform is not justified by a lower performance of their loans. The discrimination present in this online credit market has thus no economic justification, which indicates that Lending Club has an animus towards borrowers from geographic zones with a predominant African-American population.

The structure of our research paper is as follow. In section 2, we start by defining credit scoring and the different types of discrimination present on the credit market. We also introduce the emergence of FinTech lending companies and how they add elements of big data in their scoring process. Then, we discuss the advantages and disadvantages of this use of big data for credit scoring. In Section 3, we present Lending Club and peer-to-peer lending in general. We then describe the content of the data about loan applicants freely accessible online on Lending Club's platform. We explain our methodology for the analysis of geographic discrimination in the online peer-to-peer lending platform in Section 4, in which we also display our empirical results. In Section 5, we discuss our main findings and we conclude in Section 6.

2 Literature Background

2.1 Credit Scoring & Discrimination

Lenders consider three main factors to assess how risky a borrower might be: his credit history, his profile and the type of loan he is applying for. They also look at the borrower's credit score, which already incorporates some of these risk factors (Lending Club, n.d.). Credit scoring is a statistical technique used to analyze the credit risk of a loan applicant by forecasting the probability that he becomes delinquent or that he defaults on his loan (Mester, 1997). These regressions use variables like the applicant's outstanding debt, monthly income, credit utilization rate, previous records of defaults, etc. Banks offer lower principal amount and charge higher interest rates for a lower credit score, meaning a higher credit risk. There are several methods of calculating a credit score. Currently, a widely used credit scoring system in the USA is the Fair Isaac Corporation's credit scoring system (i.e. FICO score), it usually ranges from 300 to 850 (FICO, n.d.)[1]. This algorithm, founded in 1956, opened the door to millions of new customers since it was color blind and looked exclusively at the borrower's finance (O'Neil, 2016). Indeed, a credit score brings transparency and increases objectivity in the lending process (Mester, 1997). Furthermore, credit scoring models save money to lenders and borrowers by reducing the time needed for loan approval (Lawson, 1995).

[1] Because of the proprietary nature of the FICO score, the exact formula used to compute the score is not revealed. However, Investopedia breaks down the composition of the score into five major variables that are payment history, amount owed, length of credit history, new credit and type of credit used. This credit score considers only information from the credit report, this latter being written out by credit bureaus. The three major credit reporting bureaus in the USA are Equifax, Experian and TransUnion. This information is available online at: https://www.investopedia.com/ask/answers/05/creditscorecalculation.asp [Accessed 27 Jun. 2018].

However, the traditional credit market is not free from discrimination, and protected groups[2] such as African-Americans are usually offered worse terms and conditions (Shafer & Ladd, 1981; Black & Schweitzer, 1985; and Turner et al., 2002).[3] In the mortgage credit market, studies have shown that African-Americans and Hispanics loans applicants have much lower chances to be accepted for a loan than whites with similar credit characteristics (Browne & Tootell, 1995; Han, 2004; and Munnell & Tootell, 1996). Such racial discrimination is also seen in small business credit markets (Blanchflower, Levine & Zimmerman, 2003; and Cavalluzzo & Wolken, 2005).

Before proceeding any further, the term *discrimination* and its different types need to be properly defined. In the article from Hacker & Petkova (2017), a formal definition from the European Court of Justice specifies that "discrimination consists solely in the application of different rules to comparable situations or in the application of the same rule to differing situations". To determine if discrimination prevails in the market when individuals are treated differently, the question is then whether situations are comparable or not. In the economic literature, theories of discrimination are commonly classified into *statistical* discrimination (Phelps, 1972; Aigner & Cain, 1977; and Lundberg & Starz, 1983) or *taste-based* discrimination (Becker, 1957). Statistical discrimination occurs for business-related and maximization purposes, where the decision maker has no animus towards the discriminated group, in opposition to taste-based discrimination that has no economic justification.

From a legal perspective, discrimination is defined by Civil Rights law[4] under two main theories: *disparate treatment* and *disparate impact* (McLaughlin & Levy, 2014). On the credit market, disparate treatment arises when loan officer decisions vary depending on ethnicity or other prohibited characteristics (e.g. gender), either directly incorporated in the scoring algorithm, or through the lender's personnel biases. This form of discrimination is prohibited by the Fair Housing Act of 1968 (FHA) and the Equal Credit Opportunity Act of 1974 (ECOA) stating that any decision affecting the transaction's term is forbidden if taken "on the basis of a borrower/applicant's race, religion, national origin, gender, familial status, or handicap" (Bartlett et. Al, 2017). However, disparate treatment claims have been extended to the use of geography as an indicator variable in

[2] The US Supreme Court defined five categories of protected groups: race, national origin, alienage, sex, and non-marital parentage (Yoshino, 2011).
[3] For more literature about racial, gender and other minority groups discrimination in consumer and labor markets, see Pager & Shepherd (2008), Bertrand & Mullainathan (2004), and Ayers & Siegelman (1995).
[4] Title VII of the Civil Rights Act of 1964.

the scoring procedure, as shown in the article from Gano (2017), where geographic zones were denied access to credit because of their ethnic composition.

Concerning disparate impact, this form of discrimination occurs when a business's practice, neutral at first sight, results in the exclusion of a protected group. For instance, if an employer decides to hire only tall people, there will be an adverse impact on women since these latter are usually shorter than men, which results in less woman being hired than men (McLaughlin & Levy, 2014). In the loan market, an example of disparate impact would be a scoring algorithm that turns out to be a proxy for race, excluding only racial minorities, even if the resulting correlation has no intent of discrimination a priori (Zarsky, 2014; and Citron & Pasquale, 2014). However, statistical discrimination under the disparate impact theory is admissible for *legitimate business necessity*, meaning that the variables used in the scoring algorithm of the previous example are strongly correlated with credit risk-related factors (Bartlett et al., 2017). Hence, if some variables in a scoring model generate a disparate impact but turn out to be sufficiently related to credit risk, it might be admissible under the ECOA. Nevertheless, these variables cannot be used for other reasons, such as maximizing profit (McLaughlin & Levy, 2014).[5] Although a plaintiff can file a complaint under the FHA or ECOA for disparate impact, its ability to settle the lawsuit is hampered by its obligation to prove that "no legitimate business necessity mandated the use of a sorting mechanism that discriminated" (Bartlett, 2017).

2.2 Big Data & FinTech Lending

The emergence of financial technology (i.e. FinTech) companies expanded the landscape of credit scoring through the use of big data[6]. These FinTechs compete with existing financial institutions by using innovative technologies and non-traditional data sources such as mobile data, web browsing and social networks (Lenddo, 2016). In this respect, big data increases the ability to rank and rate individuals by feeding more complex algorithms used to compute credit risk (Citron & Pasquale, 2014). Different studies have shown how big data can predict personality traits (Golbeck et al., 2011; Rao et al. 2010; Schwartz et al., 2013), which in turn can be used as predictive variables for assessing someone's probability of default. In a study from Bachrach et al. (2012), it is shown

[5] For more examples on the matter, see Jolls (2001), Ayres (1991) and Gunter (2000).

[6] Gartner defines big data as "high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation". This information is available online at: https://www.gartner.com/it-glossary/big-data/ [Accessed 27 Jun. 2018].

how someone's personality traits can be predicted based on its Facebook profile information and its user friends' actions. This usage of social media can thus be used as signal for future behaviors, which in turn might determine one's potential credit risk (Brill, 2012).

Big data enables FinTech loan companies to improve the predictive power of their credit scoring model by increasing the number of data points in their algorithms, which significantly reduces costs in the underwriting process (McLaughlin & Levy, 2014; and Bartlett et al., 2017). As an example, Lenddo, a FinTech company founded in 2011 assisting banks in credit scoring by adding non-traditional data[7] into their scorecard, claims to reduce the default rate for loan applicants by 12% in comparison to traditional credit scoring models. By enhancing the predictive power that someone will actually engage in the behavior, big data scoring decreases the number of persons becoming delinquent or defaulting on their loan, which in turn helps banks to maximize their profit (Jennings, 2016).

In addition to the increased accuracy of credit scoring models, by knowing loan applicants better, big data scoring allows traditionally underserved community to have access to credit. Indeed, a critique of traditional credit scoring techniques is that populations with few or no reported credit history are attributed low credit scores or are even denied access to credit whereas they have never engaged in risky behavior (McLaughlin & Levy, 2014). To come back to the example of Lenddo, when launched, its main objective was to enhance the emerging middle class in developing countries by making these populations access to micro loans (Lenddo, n.d.). As a result, Lenddo achieved a 15% higher approval rate than traditional credit scoring models for a same set of applicants (Lenddo, 2016). Hence, big data in scoring models can be used in microfinance[8] to extend credit to people with no credit history but low probability of default.

Besides the higher efficiency of scoring procedures, Zarsky stipulates in its article "Understanding Discrimination in the Scored Society" from 2014 that more automated algorithms for credit scoring should be promoted for another reason. Scoring models with more predictive power through innovative procedures and wider data sources can limit discrimination on the credit market, particularly those concerning race. Other studies follow that path, stating that errors of

[7] Some of the non-traditional data sources used by Lenddo are: Social Networks, Mobile Data, Browser Data, Telecom Data, E-Commerce Transaction Data, Psychometric Data, to name a few. This information is available online at: https://www.lenddo.com/pdfs/Lenddo-Scoring-Factsheet-201608.pdf [Accessed 28 Mar. 2017]

[8] In their article "Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China", Chen, Li & Lai define microfinance as "a development tool that could provide vast number of the poor, especially women, with sustainable financial services to support their livelihood".

human judgment and bias, intentional or not, can be alleviated through more automation (Meadow & Sunstein 2001; and Zarsky, 2016). As mentioned previously, discrimination against minority groups such as African-Americans is still present on the credit market. In that respect, a more standardized process using big data for credit scoring instead of human discretion would ensure that lenders assessing the credit risk of an individual rely on neutral factors instead of their human bias, hence reducing discrimination under the disparate treatment theory. In their paper "Consumer Lending Discrimination in the FinTech Era" from 2017, Bartlett et al. compared the level of racial discrimination between traditional and FinTech lenders, with a focus on the mortgage market. Overall, they found that African-American and Hispanic loan applicants have 2% more probabilities to be rejected for their mortgage application than others. Moreover, in loan pricing, they found that this minority group pays a higher interest rate of 0.18%. However, their results indicate that the level of illegitimate discrimination present in loan approval and pricing is half as much for FinTech lenders as for traditional ones, suggesting that new FinTech algorithms for credit scoring may prove to be fairer for African-American and Hispanic borrowers.

However, big data for credit scoring also bears some risks. A critique of the credit scoring system in general is that it produces arbitrary results, as shown is a study from Carter et al. (2006), where almost 30% of loan applicants had credit scores differing by more than 50 points from one credit bureau to the other. [9] Despite this issue, traditional credit scoring systems remain somewhat transparent. For instance, an individual is offered basic instructions on how to boost its score on FICO's website (O'Neil, 2016), and is given a general understanding of how this score is computed and the relative weight of the different variables used (McLaughlin & Levy, 2014). Moreover, the industry is regulated by the Fair Credit Reporting Act (FCRA) of 1970 that gives consumers of credit bureaus the right to access their credit records containing all the information that feed the scoring algorithm, which can then ask for adjustments in case mistakes were to be found (Citron & Pasquale, 2014). With the rise of big data, individuals are given an *e-score* computed from a multitude of non-traditional data sources, from their online consumption habit to their zip codes, leaving them confounded by how their behavior may affect these big credit scores. Indeed, these individuals do not know what data is used and how they are translated into these e-scores. In addition to the impossibility for them to adjust their behavior to increase their creditworthiness,

[9] Although FICO scores stay widely used on the credit market, credit bureaus have built their own credit scoring models (Citron & Pasquale, 2014).

individuals have no way to fix potential errors that would result from these opaque algorithms, because of their unregulated and proprietary aspect (O'Neil, 2016).

Another danger from the use of big data in scoring procedures concerns its discriminatory potential. As discussed above, the use of big data for credit scoring could prevent discrimination on the credit market by increasing neutrality in the scoring process[10]. However, big data might also bring a whole new kind of proxies for protected minorities (Zarsky, 2014), which in turn can lead to disparate impact if the resulting proxy of the scoring procedure unintentionally excludes a protected group of the population. For instance, ZestFinance, a FinTech lending company founded in 2009, uses unusual data about their loan applicants to calculate risk, such as whether they use correct spelling when filling their online application form. The problem is that spelling mistakes is a signal of low education, which in turn may be correlated to ethnicity (O'Neil, 2016).

However, disparate impact could also be the result of an intentional exclusion of the protected minorities. The wave of proxies made available by big data can indeed be used by lenders to intentionally circumvent the law in a hidden way. The big data era has made lending companies able to find out some characteristics prohibited from considering, such as ethnicity (McLaughlin & Levy, 2014). As a result, it is possible for lenders to offer access to credit only to certain population (Ksherti, 2014). If we take the case of ethnicity, researches have shown that African-Americans and Latinos use the Internet differently (Fox, 2013; Tremoglie, 2013; and Hoenig, 2013). For instance, these latter are more likely to do their banking using a smartphone rather than a traditional laptop. Another study from Kosinski, Stillwell and Graepe (2013) stipulates that Facebook Likes can predict ethnicity with accuracy. Moreover, IP addresses collected via the Internet can predict someone's zip code that can then be used as a proxy for ethnicity (McLaughlin & Levy, 2014). These digital trails can thus be used for discriminatory practices, but in a way that is unnoticeable to the public, where lending companies mask the discrimination by using neutral on their face characteristics that correlate with discriminatory attributes (Petkova, 2017). Furthermore, such practices would be impossible to discover since big data scores use proprietary algorithms (McLaughlin & Levy, 2014).

The literature so far introduced big data for credit scoring as a mean to fight discrimination by limiting disparate treatment through more automation and less reliance on human decision, or

[10] Albeit some discrimination could still remain because of the *wrong of the past*, meaning that the data mined, resulting from past intentional discrimination, is itself responsible for present encountered discrimination (Barocas & Selbst, 2016)

at the contrary as a tool to strengthen discrimination by giving more possibilities for disparate impact on the credit market. Whether FinTech lending companies resolve this long-lasting issue of discrimination or bring new weapons for discriminatory practices remains an open question.

3 Lending Club: Data Analysis

3.1 Peer-to-Peer Lending

In order to add a contribution to the existing debate about the extent to which the use of big data for credit scoring would either facilitate or hinder exclusion of minorities on the credit market, we analyze discrimination in a new kind of online credit marketplace, peer-to-peer lending. Indeed, peer-to-peer lending platforms go beyond traditional factors used for underwriting loans such as FICO score, and build their own scorecard based on traditional and non-traditional sources of data (Prime Meridian Income Fund, 2015). Particularly, we consider the case of Lending Club, one of the two leaders in the US peer-to-peer lending market along with Prosper, whose data are made available online on their website[11].

Peer-to-peer lending represents an alternative credit marketplace that enables borrowers and lenders to enter into a credit agreement while avoiding usual intermediaries such as banks (Pope & Sydnor, 2011). What makes this new type of lending attractive for borrowers is that it cuts out part of the *middle man*, meaning lower overhead costs since the loan applicant does not have to undergo the usual underwriting process of traditional banks. Moreover, peer-to-peer lending enables individual lenders to diversify their lending portfolio, the same way as traditional banks do, by investing in any type of loans, from personal loans to mortgage refinancing loans (Patoka, 2018).

Regarding Lending Club, this online platform was launched in 2007 and is currently the largest worldwide peer-to-peer lending platform where borrowers can subscribe to a loan from $500 up to $40,000. As in most of the online peer-to-peer lending platforms, Lending Club requires first loan applicants and lenders to register on the lending platform's website by providing some personal information such as their identity, bank account number and address. Regarding those who register to become borrower, Lending Club grades them with a *Model Rank*[12] that uses their

[11] Cf. infra in the subsection "Data".
[12] The Model Rank ranges from 1 to 25, with 1 being the highest rank.

FICO score provided by credit bureaus, along with some other credit attributes[13], and an internally developed algorithm leveraging online data from the loan applicant (Garret, 2017). Then, the applicant has to specify the amount he wants to borrow and the repayment period, so that these information, along with its Model Rank, can be used to assign a *Loan Grade*[14] that determines the interest rate that will be charged on the loan (Patoka, 2018).[15] Investors can then choose in which loans to invest, in the form of notes[16], which results in their money being transferred to the borrower's bank account in exchange of monthly repayments of principal and interest[17]. Several investors can invest in the same loan. Moreover, since a partnership has been formed with FOLIO Investing, Lending Club allows its customers to trade their notes on a secondary market, meaning more liquidity for investors. However, they do not get paid if the borrower defaults on his loan.[18] A default from the borrower results in a decrease of his credit grade. Furthermore, the default record is also sent to credit bureaus so that his FICO score is impacted as well (Chen, Li & Lai, 2016).

In the end, this online process using technology to bring down costs allows Lending Club to "pass the savings back in the form of lower rates for borrowers and solid returns for investors" (Lending Club, n.d.). However, is this disruptive technology free from discrimination of protected groups, in particular towards African-Americans? It is the question we now attempt to answer regarding some demographic data that are made available by Lending Club in its policy of transparency.

3.2 Data

On the website of Lending Club, anyone can freely download data from funded loans and rejected applications.[19] These data are available since the company was founded, giving us a large amount of valued information about loans applicants and the condition to which they are awarded a loan. Funded loans from the first data set can have different status, six in total. Once funded, a loan can

[13] Some of these attributes are debt-to-income ratio (DTI), credit history and credit utilization rate.
[14] The Loan Grade ranges from A to G, with A being the top-rated loans and G the worst ones. An additional Sub-Grade ranging from 1 to 5 determines the different interest rates existing inside a certain Loan Grade, with 1 being the lowest, and 5 the highest.
[15] Some additional documents from the loan applicant, such as a proof of income or source of income, may still be required in some cases to demonstrate his credibility.
[16] "Notes are assets that correspond to fractions of loans, in amounts as low as $25" (Lending Club, n.d.).
[17] Net of the 1% Lending Club service fee.
[18] The default rate of Lending Club ranges from 1.4% for best-rated loans to 9.8% for worst-rated loans.
[19] These data were downloaded online at: https://www.lendingclub.com/info/download-data.action [Accessed 28 Mar. 2017]

be *Fully Paid*, *Current*, *Late 16-30 days*, *Late 31-120 days*, *Default* or *Charged off*.[20] Loan Grades ranging from A to G are also included, along with other information about the loan such as the purpose of the loan, the funded amount, the term and the interest rate. Data about the borrower are available as well, such as the ownership status of his home, his annual income, his debt-to-income (DTI) ratio, his job position and how long he held that position (i.e. employment length). Concerning the second set of data, which are rejected applications from Lending Club itself, a more restrictive list of loans and borrowers' characteristics are given. Among these characteristics, one can find the amount of loan requested, the loan purpose, the DTI ratio of the borrower and his employment length. A non-exhaustive list of the different variables present in each data set is available in *Table 1* in appendix.[21] What is surprising is that, in both data sets, the first three digits of the borrower's zip code are given. It is this last information that we use to perform our analysis for potential discrimination in this new peer-to-peer lending marketplace. Indeed, although this demographic information provided by Lending Club is limited, the first three digits of the borrower's zip code can still be matched with the geographic repartition of the different ethnic groups of the population, which can give an indication about the probability of that borrower's ethnicity.

In our analysis, we want to show if borrowers from geographic areas with a higher proportion of African-Americans are treated differently than those from other areas, controlling first for other observable credit variables, called control variables, meaning that we take any existing correlation between the discriminatory variable and these control variables into account. Then, we want to demonstrate if, at equal credit characteristics, these borrowers from areas densely populated in African-Americans are still differently considered and, if so, in what proportion. For this second examination, we need to take into account the interactions between the discriminatory variable and the control variables.

Before we can perform such analysis, we first need to add to the database provided by Lending Club, for each borrower, the relative number of African-Americans present in his

[20] The explanation of the six loan status is as follow: Fully Paid (the loans has been entirely repaid), Current (the loan is still being paid), Late 16-30 days (the loan has not been paid for 16-30 days), Late 31-120 days (the loan has not been paid for 31-120 days), Default (the loan has not been paid for more than 120 days) and Charged Off (the loan has been in Default state for more than 30 days). This information is available online at: http://rstudio-pubs-static.s3.amazonaws.com/339711_9c8fa45f2a3144a392b405bb25ff3a05.html [Accessed 12 Jul. 2018].

[21] The full list named *DATA DICTIONARY* can be downloaded online at: https://www.lendingclub.com/info/download-data.action.

geographic zone. To do so, we use the first three digits of the borrower's zip code already present in the data set. A zip[22] code is usually composed of five digits that are used by the United States Postal Service (USPS) for delivery purpose (Curtin, n.d.). The first digit indicates the US state, with several states being regrouped in a same number. The second and third digits represent together a certain region in that state, or sometimes a large city, while the two last digits are used to identify groups of mailing addresses in that region (Kabramson, 2015). From a map of the USNaviguide (2010)[23], we can identify to what geographic area each first three digits of a zip code correspond. However, some of these areas incorporate different small towns or incomplete parts of county, which makes the demographic analysis difficult. Therefore, we restrict our analysis to large cities only, since their boundaries are clearly delimited by a unique three digits zip code. This procedure gives us a total of 201 three digits zip codes[24]. The 201 big cities corresponding to these three digits zip codes can then be further analyzed for their ethnic composition. For each of these cities, we can find on the website of the US Census Bureau[25] its demographic statistics, including the racial composition of the population. This final step enables us to associate to each borrower from large cities the percentage of African-Americans living in his three digits zip code area.

Since the ethnic composition given on the US Census Bureau's website are the result of the "American Community Survey" from 2016, we conduct our tests on Lending Club's data of the year 2016 as well. Over that period, 434,371 loans were funded from which we keep 113,126 loans for our analysis, for which the first three digits of the borrower's zip codes correspond to clearly defined areas. This sample represents 26% of the original set of funded loans. The average size of a loan on that sample is $14,446 while the interest rate ranges between 5% and 31% and amounts on average to 12.87%. The term length for the loans is either 36 months or 60 months with a clear majority of 36 months loans in the sample year. Most of the loans have the Current status (52.72%). Loans are mainly graded B (31.29%) and C (30.20%) with only 0.57% of G. Concerning borrowers, the median annual income is $65,000 and the median DTI ratio amounts to 17.89%. Among them, only 10.32% own their home. In the second data set, among the 4,576,653 rejected applications, we keep 1,224,091 applications, for the same reason than in the first data set, representing this time almost 27% of the original data. This summary statistics is available in *Table 2* in appendix.

[22] The word ZIP is an acronym for *Zone Improvement Plan*.
[23] This map is available online at: http://maps.huge.info/zip3.htm [Accessed 12 Jul. 2018].
[24] Out of the 929 three digits zip codes present in the USA.
[25] https://www.census.gov/quickfacts/fact/table/US/RHI225217#viewtop [Accessed 12 Jul. 2018].

For all borrowers, we assign a dummy variable that takes the value 1 for those living in areas with more than 50% of African-Americans, and 0 for the ones living in areas with 50% or less. This categorization helps us perform the statistical tests that are presented in the next section. When dividing for each of the two categories the number of borrowers accepted for a loan by the total amount of loan applicants[26], we find that 7.56% of the borrowers from areas highly populated in African-Americans have been granted a loan against 9.47% for the other borrowers. Moreover, we see that the group of borrowers located in places with a predominantly African-American community have an average interest rate of 13.12%, compared to an average of 12.84% for the other group. These results are shown in the summary statistics in *Table 2* in appendix. Of course, we cannot conclude at this stage that borrowers from geographic areas densely populated in African-Americans have less chances to be accepted for a loan and are charged higher interest rates, since this characteristic might be strongly correlated with other credit characteristics. As mentioned earlier, to analyze such differential treatment, we then need to perform more rigorous tests that take into account control variables.

4 Methodology & Empirical Results

4.1 Acceptance Rate

In this first subsection, we examine in what proportion the density of African-Americans that live in the housing area of the borrower affects his probability to be accepted by Lending Club for a loan, while controlling for the other available credit characteristics. Although the summary statistics of the previous section indicates that geographic exclusion might be present in the process of loan approval, we now perform more sophisticated statistical regressions to determine if these results hold in presence of control variables. A statistical regression is used to find estimators for the coefficients of a model explaining the relationship between a response variable and explanatory variables. In this case, we use binomial logistic regressions since the response variable is a dummy variable that takes the value 1 when the applicant is accepted and 0 if he is rejected.[27] Indeed, "logistic regression is a method for fitting a regression curve, $y = f(x)$, when y is a categorical variable" (Michy, 2015). In total, we perform three logistic regressions on three different models.

[26] To find the total amount of loan applicants, we sum the number of funded loans and the number of rejected applications.

[27] Our logistic regression framework has the form: $\ln(odds_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + X_i\beta$, where π_i is the probability that the loan *i* is accepted and X_i is the matrix of explanatory variables.

The first model considers only control variables as explanatory variables, the second one adds the density of African-Americans that are present in the borrower's living area in the form of a dummy variable, as explained in the previous section, while the third model takes into account interactions between this latter variable and the former control variables. For simplicity reasons, we name this dummy variable *dummy density*. The results of these regressions are shown in *Table 3* in appendix.

The control variables used in our different models consist in the amount of requested loan by the applicant[28], his DTI ratio and his employment length, which are the only variables, among the limited number of financial variables present in the data set of rejected loans, that we find relevant to explain the acceptance probability. The results of the logistic regression on the first model considering only these control variables confirm that the acceptance probability of an application is influenced by the three variables we keep for our analysis. The effect of the amount requested by a borrower (β = 0.00000673, $p < 0.001$) and his employment length (β = 0.402, $p < 0.001$) have both significant, positive, effects on his acceptance probability by Lending Club. It is somewhat surprising that the amount of loan requested has a positive effect on the acceptance probability of the borrower, although this impact is quite weak. Concerning the DTI ratio (β = -0.509, $p < 0.001$), its effect on the borrower acceptance probability is significantly negative.

When we add the dummy density variable, the results of the regression on this second model indicate that the density of African-Americans (β = -0.228, $p < 0.001$) has a significant, negative, impact on the acceptance probability of the borrower. These findings suggest that borrowers from areas densely populated in African-Americans are less likely to be accepted for a loan on the online peer-to-peer platform. More precisely, the log odds ratio coefficient of -0.228, resulting in an odds ratio coefficient of 0.80 for the dummy density variable[29], indicates that the odds of being accepted is 20% lower for the borrowers coming from regions mainly populated in African-Americans, after controlling for the observable financial variables we keep for our analysis.

The last model includes interactions between the dummy density variable and the control variables from which results display positive and significant coefficients for the three interactions, between dummy density and amount requested (β = 0.00000264, $p < 0.001$), between dummy density and DTI ratio (β = 0.153, $p < 0.01$) and between dummy density and employment length (β = 0.0229, $p < 0.001$). This means that, although the amount requested by the borrower and his

[28] In the data set of accepted loan, this variable is simply the amount of funded loan.

[29] In a logit model, β's are the log odds coefficients, meaning that the odds coefficients are e^{β}'s.

employment length have both significantly positive effects on its acceptance probability, these effects are even higher for the borrowers living in areas with a main density of African-Americans. It also means that, although the DTI ratio of a borrower has a significant, negative, effect on its acceptance probability, this negative effect is weaker for those same borrowers from geographic areas with a high population of African-Americans. However, in this third model, the results of the logistic regression show that the dummy density variable alone (β = -0.391, $p < 0.001$) has a higher significantly negative effect on the borrower's acceptance probability than in the previous model. These findings illustrate that borrowers from areas densely populated in African-Americans have even less probability to be accepted for a loan compared to the second model, but this probability to be accepted rises quicker than for the other group of borrowers when the employment length increases.[30] Nevertheless, this same probability, on the contrary, increases more slowly than for the other borrowers when the DTI ratio decreases.

4.2 Interest Rate

We now study the impact of the density of African-Americans present in the borrower's living area on the interest rate he must pay on his loans, once his application is accepted by Lending Club, while still controlling for observable financial characteristics. Once again, the summary statistics reveals that geographic inequalities may exist in the way Lending Club charges interest rates to borrower, but more complex statistical tools are needed to determine if these results stay valid in presence of control variables. In this case, we use simple Ordinary Least Squares (OLS) regressions to find estimators for the coefficient of the different models we use to explain the relationship between interest rate and explanatory variables. The response variable, interest rate, is no more a categorical variable so multiple linear regressions can be used.[31] We perform three OLS regressions on three different models. As in the previous subsection, the explanatory variables of the first model are only composed of control variables, while the second model adds the dummy density variable. Lastly, the third model incorporates into the explanatory variables the interactions between the dummy density variable and the control variables. The results of these regressions are shown in *Table 4* in appendix.

[30] We do not mention the interaction between the amount of loan requested and the dummy density variable since its impact is very small on the acceptance probability of the borrower.

[31] Our OLS regression framework is as follow: $Y_i = \alpha + X_i\beta$, where Y_i represents the interest rate on the loan i and X_i is the matrix of explanatory variables.

This time, control variables represent fourteen credit variables that we choose among the hundred variables available in the data set of accepted loans. We limit ourselves to this relatively small number of variables because we find they already explain quite acceptably, using linear regressions, how interest rates are charged to borrowers.[32] Adding more control variables to the regression would only increase by very little the accuracy of our model, while making it much more complicated. The variables we keep for our analysis are the amount of funded loan, the term of the loan, the employment length of the borrower, his ownership status, his annual income, his DTI ratio, his number of delinquencies over the past two years prior his application for a loan, the length of his credit history, his inquiries in the last six months, his number of opened credit lines, his number of public records, his revolving balance, his revolving utilization and his total number of credit lines.

Among these control variables, the funded amount ($\beta = 0.000000428$, $p < 0.001$), the term ($\beta = 0.0424$, $p < 0.001$), the home ownership status ($\beta = 0.00499$, $p < 0.001$), the DTI ratio ($\beta = 0.0968$, $p < 0.001$), the number of past delinquencies ($\beta = 0.00288$, $p < 0.001$), the number of inquiries ($\beta = 0.0131$, $p < 0.001$), the number of opened credit lines ($\beta = 0.000238$, $p < 0.001$), the number of public records ($\beta = 0.00541$, $p < 0.001$) and the amount of revolving utilization ($\beta = 0.0362$, $p < 0.001$) have all significant, positive, effect on the final interest rate charged to the borrower, although the impact of funded amount is quite small. What is surprising is that the home ownership status has a positive impact, meaning that borrowers who own their home pay a higher final interest rate. Concerning the employment length ($\beta = -0.000214$, $p < 0.001$), the annual income ($\beta = -0.0000000337$, $p < 0.001$), the length of credit history ($\beta = -0.000851$, $p < 0.001$), the revolving balance ($\beta = -0.000000183$, $p < 0.001$) and the total number of credit lines ($\beta = -0.000491$, $p < 0.001$), their effects are all significantly negative on the interest rate charged to borrowers, although the impacts of the annual income and the revolving balance are very low.

Concerning the second model, in which we add the dummy density variable, the results of the linear regression indicate that the density of African-Americans has a significantly positive impact ($\beta = 0.00185$, $p < 0.001$) on the final interest rate. This means that a borrower that lives in an area mainly populated in African-Americans will have an interest rate increased by almost 20

[32] We find an Adjusted-R^2 of 0.302 when using the fourteen control variables we keep to build our model, which means that more than 30% of the total variance in the data is explained by our model.

basis points compared to other borrowers, after controlling for the observable credit characteristics we keep for our analysis.

When we add interactions between the dummy density variable and control variables, the linear regression on this last model return only four interactions that have significative effects on the interest rate. Three of these interactions, between dummy density and funded amount (β = 0.000000130, $p < 0.05$), between dummy density and home ownership status (β = 0.00253, $p < 0.05$) and between dummy density and DTI ratio (β = 0.0176, $p < 0.001$) have significant, positive coefficients, although those of the first two interactions are only significative at the 5% level. Concerning the interaction between dummy density and annual income (β = -0.0000000313, $p < 0.001$), its coefficient is significantly negative. Interestingly, the impact of the dummy density variable alone (β = -0.00151) on the interest rate charged to the borrower becomes insignificant. Such findings imply that the effect of the density of African-Americans alone on the interest rate is entirely moderated by the other observable credit variables we add in this third model. In that model, borrowers from areas densely populated in African-Americans are not initially charged a higher interest rate. However, at equal DTI ratio with the other borrowers, their interest rates are higher.[33]

4.3 Default Rate

We also investigate the relationship between the density of African-Americans that live in the borrower's housing area and the loan performance of that same borrower. The proxy used for loan performance is the default rate, computed by dividing the number of loans in Default or Charged Off status by the total number of loans. From the data set of accepted loans, we can determine the default rates of both borrowers' categories. Such calculations indicate that the group of borrowers from regions densely populated in African-Americans have a default rate of 12.30% compared to 11.64% for the other group of borrowers, as shown in the summary statistics in *Table 2* in appendix. Although the default rate is slightly higher for the first category of borrowers, no hasty conclusion can be drawn since, once again, this density characteristics may be highly correlated with other financial characteristics.

[33] We do not mention the interactions between the dummy density variable and the funded amount, the ownership status or the annual income, since their impacts on the final interest rate are either very small or not as significative as for the DTI ratio.

In order to determine if differences in default rates exist between the two categories of borrowers, taking this time into consideration control variables, we make use of the Cox proportional-hazards model (Cox, 1972). This model investigates the link between the survival time of a loan, *t*, which is the time before the loan defaults, and explanatory variables. In the Cox proportional-hazards model, the default event is treated as a positive random variable and the response variable, default rate, is the hazard function, $\lambda(t)$.[34] As a large number of loans in our sample had not matured yet, we estimate loan default rates by the hazard function that computes the default rate of a loan at time *t*, conditionally to his "survival" until time *t*, with *t* being computed as the number of days that separate the inception of the loan and its default, if any. Thus, for our analysis, we exclude from the sample all the loans that have the Fully Paid status, which represent about a third of the loans, since these latter cannot default anymore.

We perform two Cox statistical regressions on two models to find estimators for the coefficients of these different models. In the first model, the default rate of a loan is only predicted by control variables, that are the same than those used in the previous models explaining the interest rate. In the second model, we add the dummy density variable to the explanatory variables, as we did in the previous subsections. The results of the first regression indicate that, excepted for the home ownership status and the total number of opened credit lines, all other control variables have significative effects on the default rate, and thus on the length of the loan survival time.[35] However, the results of the second regression on the model that includes the dummy density variable show no significant impact from that variable on the default rate of a loan. These findings, shown in *Table 5*, illustrate that, controlling for observables credit characteristics, we cannot conclude that the density of African-Americans present in a borrower's living area has a particular impact on that borrower's loan performance.

[34] Our hazard function framework is as follow $\lambda_i(t) = \lambda_0(t) * e^{X_i\beta}$, where the response variable, $\lambda_i(t)$, is the default rate of the loan *i*, $\lambda_0(t)$ is the baseline hazard rate, which gives the value of the hazard when X_i, the matrix of explanatory variables, is equal to zero.

[35] A positive value of a variable's coefficient indicates that an increase in the variable increases the occurrence of the event default, which decreases the length of the loan's survival time. A similar reasoning can be applied in case of a negative value of the variable's coefficient (STHDA, n.d.).

5 Discussion

5.1 Major Findings

Our tests measure the impact of the borrower's location on his acceptance probability for a loan and, if accepted, on the interest rate that is charged on that loan by Lending Club, the biggest US online peer-to-peer platform. Our results indicate that borrowers from areas mainly populated in African-Americans, that is to say with more than 50% of African-Americans, have a more restricted access to credit in the online peer-to-peer market than the other group of borrowers, after controlling for other observable credit characteristics. In particular, our findings show that these borrowers have 20% less probability to be accepted for a loan and an interest rate that is 20 basis points higher compared to borrowers from regions with less than 50% of African-Americans. Moreover, when including in our model interactions between control variables and the categorical variable that indicates the density of African-Americans, we find that the probability for those borrowers to be accepted for a loan is even lower. However, this probability rises quicker than for the other borrowers' group when credit variables such as employment length increase. Still with regard to this latter model that incorporates the different interactions, we see that borrowers from areas highly populated in African-Americans are not initially charged a higher interest rate. Yet, at equal financial characteristics, such as the DTI ratio, those borrowers must pay higher interest rates compared to the other borrowers.

These results suggest that discrimination towards borrowers from regions predominantly populated in African-Americans exists in the peer-to-peer lending market. However, this discrimination cannot be classified as statistical, in the sense of profit-based. Indeed, the default rate, used as proxy for the borrower's loan performance, is not significantly impacted by the density of African-Americans living in the borrower's housing area. Thus, the lower acceptance rate and higher interest rate of the discriminated borrowers are not statistically justified by a higher default rate. However, even if the density characteristic of a borrower would have a significant, positive, effect on his default rate, meaning that borrowers from areas highly populated in African-Americans have higher default rates, after controlling for other observable variables, this result would not be sufficient to talk about accurate statistical discrimination. We would still have to compare this default rate with the extra interest rate borrowers have to pay on their loan. Since in our tests the dummy density variable is not even significative, we do not have to make any comparison with interest rates and we can thus directly talk about taste-based discrimination

towards borrowers from areas with a higher proportion of African-Americans. However, these findings are the results of tests made on a limited set of variables provided by Lending Club. As discussed further in the next subsection, these findings may not hold if more or other credit variables were to be used.

Now that we highlighted the interesting relationship between the ethnic composition of the living area of a borrower and his credit accessibility, what could explain that borrowers from regions densely populated in African-Americans are more financially constrained? One could argue that, if more than 50% of the population in a geographic zone belongs to the African-American community, the probability is greater that someone living in that area is an African-American rather than any other ethnicity. Based on this assumption, and on the fact that racial discrimination is still prevalent on the credit market, one could assume that Lending Club is using the borrower's location as a proxy for his race, which enables the lending peer-to-peer platform to discriminate African-American borrowers accordingly. Especially that, unlike us, Lending Club is provided with the full address of the borrower, and not just the first three digits of his zip code, which could allow the online platform to make even more precise assumptions about the borrower's ethnicity. It is even possible that Lending Club is using more proxies for race than the borrower's location through the use of non-traditional data on their applicants, which would enable the online peer-to-peer platform to refine their discrimination while still being unnoticed to the public.

The discrimination observed on the online peer-to-peer lending market would not be considered as disparate treatment if the density characteristic is not directly incorporated in the scoring algorithm of Lending Club, used to decide what borrower to accept for a loan and on what terms. In such a case, the resulting discrimination would be treated as disparate impact since, although not explicitly discriminating borrowers from regions highly populated in African-Americans, the outcome of this scoring algorithm would exclude the minority group from the credit market more than the other loan applicants. Moreover, this discrimination could be totally unintended by Lending Club but would be the result of a correlation between the specific ethnic density characteristic of the borrower's living area and other credit variables used in the scoring algorithm that are not made available to us by the online platform.[36] Because of the proprietary nature of the Lending Club's scoring algorithm, none of these hypotheses can be verified.

[36] In such a case, the discrimination would be statistical, meaning that our models are not accurate enough to detect the statistical discrimination even though it is the case in practice.

5.2 Limitations

The proprietary character of the Lending Club's scoring algorithm that we just discussed is also the central point of our limitations. Even though we control for several credit characteristics that we have identified as the most relevant ones among all the available data, other important variables are not taken into account in our regression models. Their absence in this research paper is not voluntary but is the result of Lending Club's intellectual property. Indeed, for obvious reasons of competition, Lending Club does not publicly provide the formula of its scoring algorithm, nor the full set of data used in the scoring procedure. As a result, our findings encounter a problem of omitted variables. For instance, there are data that are not made available to us, but that we know for sure are used in the Lending Club's Model Rank, such as the borrower's FICO score and non-traditional online data. Moreover, even if these data were accessible, we would not know how to translate them into this Lending Club score. Finally, since the variable incorporating the ethnic density of the borrower's living area may be highly correlated with these unobservable characteristics kept secret by Lending Club, our tests might give different results if they were applied on the full set of data, using the right scoring formula. However, even if these different results would demonstrate that no taste-based discrimination is present in the Lending Club's scoring procedure, a statistical discrimination would still remain given the results of the summary statistics.

Another limitation from our study is related to its methodology. We use the first three digits of the borrower's zip code to determine the percentage of African-Americans present in the borrower's housing area. As we said earlier, by using the full address of the borrower, one could make an even more precise analysis of the ethnic composition of his living place. However, this increase in accuracy could also modify our results. Indeed, one can fairly assume that the repartition of the population is not homogeneous within the large cities we keep for our demographic analysis. For instance, although a large city whose boundaries are clearly delimited by a unique three digits zip code is composed of over 70% of African-Americans, a borrower from that city could still live in the Hispanic or Chinese neighborhood, where the percentage of African-Americans is well below the average 70%.

5.3 Recommendations

Since further research incorporating more variables into the different models that attempt to explain how borrowers' loan applications are handled by Lending Club is not possible due to the secrecy

nature of these new big data algorithms, we end this section by proposing some recommendations to detect the presence of discrimination, and how to manage it if any, on this online peer-to-peer lending platform and for FinTech lending companies in general.

Transparency measures should be applied so that each loan applicant would have access to their data, as well as the way these data are used in the scoring algorithm and the final result of the credit scoring process. This way, individuals would be able to identify any resulting discrimination from their predictive e-scores and would have the opportunity to challenge them (Citron & Pasquale, 2014). For instance, Article 12 of the EU General Data Protection Directive (GDPR) requires data companies to provide subjects of rating with a detailed record of data-based decisions that affect their rating (Zarsky, 2016). More directives in line with this transparency approach should be enacted.

However, this idea of increased transparency in big data algorithmic decision making has two main disadvantages. First, these transparency requirements would reveal to the public the secret recipe of the lending company's big scoring algorithm, which represents the most valuable part of the business model of the company. As a result, innovation would be endangered in the credit industry. Second, more transparency would allow individuals to identify the behavioral factors of credit risk used in the scoring algorithm and their different weights, which in turn can lead some of these scored individuals to send false or amplified signals about the most important indicators of creditworthiness. In other word, this would allow borrowers to *game* these new algorithms using big data for credit scoring (Hacker & Petkova, 2017).

A solution to these concerns raised by more transparency towards the public at large would be to limit disclosure of the proprietary big credit scoring algorithms to impartial experts. By licensing and auditing FinTech lending companies that use new big data scoring systems, these independent third parties would limit disparate treatment by ensuring no illegitimate factors such as race are used in the algorithm, while protecting the intellectual property of these big data scoring companies (Citron & Pasquale, 2014). Moreover, in order to mitigate disparate impact as well, these neutral experts would adjust or even remove from the scoring mechanism factors contributing to the creation of a disparate impact. Nevertheless, this last step might diminish the predictive power of these new scoring algorithms since important predictors of creditworthiness may be removed from it. These adjustments of scoring factors made in the algorithm to limit disparate

impact against minorities could even lead to a new form of discrimination, this time towards the majority group (Zarsky, 2016). In the end, there is no solution that is not a zero-sum game.

6 Conclusions

The use of big data for credit scoring by FinTech lending companies enhance the precision of scoring algorithms since this new kind of lenders better assess borrowers' repayment ability by adding non-traditional data such as browser data, social network and mobile data to traditional ones in the scoring procedure. This higher accuracy of big data scoring reduces the default rate of borrowers and increases the approval rate of loans applicants. Although the ability of big data to reach higher levels of efficiency in credit scoring is little questioned, it is a different story regarding the fairness it brings on the credit market. On the one hand, big data could lower reliance on human decision and thus reduce human bias in scoring procedures by making scoring algorithms more automated, which would limit the risk of disparate treatment on the credit market. On the other hand, big data for credit scoring could facilitate lending companies to discriminate minority groups in a concealed manner since it brings new possibilities of proxies for the protected classes of the population.

To find elements of response in this literature gap, we search for the presence of discrimination in a new credit marketplace, online peer-to-peer lending, since lending platforms operating on this market use non-traditional data to refine their credit scoring algorithms. In particular, we investigate the existence of geographic discrimination of areas predominantly populated in African-Americans in the biggest US lending platform, Lending Club, whose data about loan applicants are partly made available to the public. Since these open data incorporate the first three digits of the zip codes of loan applicants, we are able for each of them, using demographic data from the Census Bureau, to determine the proportion of African-Americans present in their living area. Then, we analyze the extent to which the ethnic density characteristic of an applicant impacts his acceptance probability for a loan and the interest rate he must pay on that loan.

The results of our tests indicate that, after controlling for other observable credit characteristics, the group of borrowers living in areas where the African-American community exceeds 50% of the population suffers from discrimination in the online peer-to-peer lending market. These borrowers have 20% less probability to be accepted for a loan and they are charged an interest rate on their loan that is 20 basis points higher compared to the other loan applicants.

Since the loans performance of these discriminated borrowers, measured by their default rates, is not inferior to the performance of the other borrowers' loans, the discrimination present in the online peer-to-peer platform is taste-based. However, it is important to keep in mind that the list of data accessible on the platform's website is limited and that more variables are used in the scoring algorithm of Lending Club. Therefore, our different models encounter a problem of omitted variables.

Even though our tests show geographic discrimination is present in the online peer-to-peer lending market, our study does not allow us to affirm that this discrimination is higher or lower in such market, where lending companies use big data for credit scoring, compared to other traditional credit markets. To answer this question in the case of Lending Club, we would have to compare our results with those of similar tests made on the same set of data, but where Lending Club would not use non-traditional online data from loan applicants in its scoring procedure. This way, if the tests performed in the second situation, where the scoring algorithm is free from big data components, would give results even more disadvantageous for the minority group, we could assume that big data for credit scoring would at least alleviate part of the discrimination. However, given that our results indicate the presence of taste-based discrimination towards borrowers from areas with a predominant African-American population in the online peer-to-peer platform, we can assert that the use of big data for credit scoring does not remove from the credit market geographic discrimination excluding racial minorities.

7 Appendix

Table 1

List of Variables used in Lending Club's Data Sets

Accepted Loans	Description
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application
addr_state	The state provided by the borrower in the loan application
annual_inc	The annual income provided by the borrower during registration
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years
emp_title	The job title supplied by the Borrower when applying for the loan
funded_amnt	The total amount committed to that loan at that point in time
funded_amnt_inv	The total amount committed by investors for that loan at that point in time
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration, our values are: RENT, OWN, MORTGAGE, OTHER
id	A unique LC assigned ID for the loan listing
initial_list_status	The initial listing status of the loan. Possible values are – W, F

Accepted Loans	Description
inq_last_6mths	The number of inquiries by creditors during the past 6 months
installment	The monthly payment owed by the borrower if the loan originates
int_rate	Interest Rate on the loan
is_inc_v	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower
loan_status	Current status of the loan
member_id	A unique LC assigned Id for the borrower member
mths_since_last_delinq	The number of months since the borrower's last delinquency
mths_since_last_major_derog	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record
next_pymnt_d	Next scheduled payment date
open_acc	The number of open credit lines in the borrower's credit file
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
policy_code	Publicly available policy_code=1, new products not publicly available policy_code=2
pub_rec	Number of derogatory public records
purpose	A category provided by the borrower for the loan request
pymnt_plan	Indicates if a payment plan has been put in place for the loan

Accepted Loans	Description
recoveries	Post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan, values are in months and can be either 36 or 60
title	The loan title provided by the borrower
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
total_rec_prncp	Principal received to date
url	URL for the LC page with listing data

Rejected Loans	Description
Zip Code	The first 3 numbers of the zip code provided by the borrower in the loan application
Amount Requested	The total amount requested by the borrower
Application Date	The date which the borrower applied
Loan Title	The loan title provided by the borrower
Debt-To-Income Ratio	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income

Rejected Loans	Description
State	The state provided by the borrower in the loan application
Employment Length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years
Policy Code	Publicly available policy_code=1, new products not publicly available policy_code=2

Table 2

Summary Statistics

Information from Funded Loans

Variables	Funded Loans
Loan Information	
Average Size	14,446
Term Length Majority	36m
Average Interest Rate	12.87%
Borrower Information	
Median Annual Income	65,000
Median DTI Ratio	17.89%
Owns a Home	10.32%
Loan Status	
Fully Paid	33.55%
Current	52.72%
Late (16-30 days)	0.35%
Late (31-120 days)	1.55%
Default	0.00%
Charged Off	11.82%
Loan Grade	
A	16.49%
B	31.29%
C	30.20%
D	13.52%
E	5.81%
F	2.12%
G	0.57%
Funded Loans	113,126
Total Applications	1,224,091

African-American Density	Acceptance Rates
> 50%	7.56%
≤ 50%	9.47%

African-American Density	Interest Rates
> 50%	13.12%
≤ 50%	12.84%

African-American Density	Default Rates
> 50%	12.30%
≤ 50%	11.64%

Table 3

Effects of Density on Acceptance Rate

	(1) Control Variables	(2) Add Density	(3) Add Interactions
Amount Requested	0.00000673***	0.00000667***	0.00000641***
	(38.14)	(37.65)	(34.20)
DTI Ratio	-0.509***	-0.508***	-0.525***
	(-28.37)	(-28.35)	(-27.52)
Employment Length	0.402***	0.402***	0.399***
	(389.65)	(389.44)	(367.38)
Dummy Density		-0.228***	-0.391***
		(-19.40)	(-16.28)
Amount Requested * Dummy Density			0.00000264***
			(4.58)
DTI Ratio * Dummy Density			0.153**
			(2.79)
Employment Length * Dummy Density			0.0229***
			(6.66)
Requested Loans	1,224,075	1,224,075	1.224,075
Pseudo R^2	0.2482	0.2487	0.2488

Logistics regressions are used; t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4

Effects of Density on Interest Rate

	(1) Control Variables	(2) Add Density	(3) Add Interactions
Funded Amount	0.000000428***	0.000000429***	0.00000042***
	(26.29)	(26.36)	(24.59)
Term	0.0424***	0.0424***	0.0424***
	(134.65)	(134.60)	(127.69)
Employment Length	-0.000214***	-0.000216***	-0.000211***
	(-6.60)	(-6.64)	(-6.16)
Home Ownership Status	0.00499***	0.00492***	0.00459***
	(12.26)	(12.08)	(10.53)
Annual Income	-0.0000000337***	-0.0000000336***	-0.0000000324***
	(-20.57)	(-20.53)	(-19.41)
DTI Ratio	0.0968***	0.0967***	0.0951***
	(69.26)	(69.22)	(65.31)
Delinquencies over the past 2y	0.00288***	0.00288***	0.00286***
	(21.93)	(21.87)	(20.56)
Length of Credit History	-0.000851***	-0.000852***	-0.000851***
	(-49.43)	(-49.48)	(-46.72)
Inquiries in the last 6m	0.0131***	0.0131***	0.013***
	(88.73)	(88.76)	(84.06)
Number of Opened Credit Lines	0.000238***	0.000237***	0.00024***
	(7.45)	(7.42)	(7.09)
Public Records	0.00541***	0.00538***	0.0054***
	(27.71)	(27.57)	(25.84)
Revolving Balance	-0.000000183***	-0.000000183***	-0.000000182***
	(-30.98)	(-30.96)	(-29.09)
Revolving Utilization	0.0362***	0.0362***	0.0364***
	(65.87)	(65.86)	(62.90)
Total Number of Credit Lines	-0.000491***	-0.000492***	-0.000499***
	(-31.58)	(-31.62)	(-30.32)
Dummy Density		0.00185***	-0.00151
		(4.46)	(-0.82)
Funded Amount * DD[37]			0.000000130*
			(2.28)
Home Ownership Status* DD			0.00253*
			(2.06)
Annual Income * DD			-0.0000000313***
			(-3.66)
DTI Ratio * DD			0.0176***

[37] DD stands for Dummy Density.

	(1) Control Variables	(2) Add Density	(3) Add Interactions
			(3.35)
Funded Loans	113,054	113,054	113,054
Adjusted-R^2	0.3020	0.3021	0.3024

OLS regressions are used; t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5

Effects of Density on Default Rate

	(1) Control Variables	(2) Add Density
Funded Amount	0.0000251***	0.0000251***
	(19.81)	(19.81)
Term	0.0951***	0.0951***
	(4.56)	(4.56)
Employment Length	-0.0107***	-0.0107***
	(-4.64)	(-4.64)
Home Ownership Status	0.0416	0.0416
	(1.44)	(1.44)
Annual Income	-0.00000451***	-0.00000451***
	(-16.53)	(-16.53)
DTI Ratio	0.340***	0.340***
	(12.69)	(12.69)
Delinquencies over the past 2y	0.0248**	0.0248**
	(2.90)	(2.90)
Length of Credit History	-0.0148***	-0.0148***
	(-11.33)	(-11.33)
Inquiries in the last 6m	0.269***	0.269***
	(30.39)	(30.39)
Number of Opened Credit Lines	0.000341	0.000341
	(0.16)	(0.16)
Public Records	0.0834***	0.0834***
	(7.16)	(7.16)
Revolving Balance	-0.00000884***	-0.00000884***
	(-10.78)	(-10.77)
Revolving Utilization	0.324***	0.324***
	(7.88)	(7.88)
Total Number of Credit Lines	0.00781***	0.00781***
	(7.26)	(7.26)
Dummy Density		-0.000667
		(-0.02)
Total Loans	74,203	74,203
Pseudo-R^2	0.029	0.029

Cox regressions are used; t-statistics in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

8 Bibliography

Aigner, D. J., & Cain, G. G. (1977). Statistical theories of discrimination in the labor market. Industrial and Labor Relations Review, 30(2), 175-87.

Ayres, I. (1991). Fair driving: Gender and race discrimination in retail car negotiations. Harvard Law Review, 817-872.

Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. The American Economic Review, 304-321.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. In Proceedings of the 4th Annual ACM Web Science Conference (pp. 24-32). ACM.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Cal. L. Rev., 104, 671.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2017). Consumer Lending Discrimination in the FinTech Era. Working paper, University of California, Berkeley

Becker, G. S. (2010). The economics of discrimination. University of Chicago press.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review, 94(4), 991-1013.

Black, H. A., & Schweitzer, R. L. (1985). A canonical analysis of mortgage lending terms: Testing for lending discrimination at a commercial bank. Urban Studies, 22(1), 13-19.

Blanchflower, D. G., Levine, P. B., & Zimmerman, D. J. (2003). Discrimination in the small-business credit market. Review of Economics and Statistics, 85(4), 930-943.

Brill, J. (2012). Big data, big issues. US FTC.

Browne, L. E., & Tootell, G. M. (1995). Mortgage lending in Boston-A Response to the Critics. New England Economic Review, 53-72.

Carter, C., Renuart, E., Saunders, M., & Wu, C. C. (2006). The Credit Card Market and Regulation: In Need of Repair. NC Banking Inst., 10, 23.

Cavalluzzo, K., & Wolken, J. (2005). Small business loan turndowns, personal wealth, and discrimination. The Journal of Business, 78(6), 2153-2178.

Chen, D., Li, X., & Lai, F. (2017). Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in China. Electronic Commerce Research, 17(4), 553-583.

Citron, D. K., & Pasquale, F. (2014). The scored society: due process for automated predictions. Wash. L. Rev., 89, 1.

Cox, D. R. (1972). Models and life-tables regression. JR Stat. Soc. Ser. B, 34, 187-220.

Curtin, A. (n.d.). Flashing Across the Country: Mr. Zip and the ZIP Code Promotional Campaign. [Online] Available at: https://postalmuseum.si.edu/zipcodecampaign/ [Accessed 12 Jul. 2018].

FICO, (n.d.). FICO at a Glance. [Online] Available at: http://www.fico.com/en/about-us#at_glance [Accessed 27 Jun. 2018].

Fox, S. (2013). 51% of U.S. adults bank online. [Online] Available at: http://www.pewinternet.org/2013/08/07/51-of-u-s-adults-bank-online/ [Accessed 1 Jul. 2018].

Gano, A. (2017). Disparate Impact and Mortgage Lending: A Beginner's Guide. U. Colo. L. Rev., 88, 1109.

Garret, O. (2017). The 4 Best P2P Lending Platforms For Investors In 2017 - Detailed Analysis. [Online] Available at: https://www.forbes.com/sites/oliviergarret/2017/01/29/the-4-best-p2p-lending-platforms-for-investors-in-2017-detailed-analysis/#6f3396a652ab [Accessed 10 Jul. 2018].

Golbeck, J., Robles, C., & Turner, K. (2011, May). Predicting personality with social media. In CHI'11 extended abstracts on human factors in computing systems (pp. 253-262). ACM.

Gunter, K. G. (2000). Computerized Credit Scoring's Effect on the Lending Industry. NC Banking Inst., 4, 443.

Hacker, P., & Petkova, B. (2017). Reining in the big promise of big data: transparency, inequality, and new regulatory frontiers. Nw. J. Tech. & Intell. Prop., 15, i.

Han, S. (2004). Discrimination in lending: Theory and evidence. The Journal of Real Estate Finance and Economics, 29(1), 5-46.

Hoenig, C. (2013). Survey: Mobile Banking Most Common Among Blacks, Latinos. [Online] Available at: www.diversityinc.com/news/survey-mobile-banking-most-common-amongblacks-latinos [Accessed 1 Jul. 2018].

Jennings, S. (2016). Expert Interview Series: Erki Kert of Big Data Scoring - Syncsort Blog. [Online] Available at: http://blog.syncsort.com/2016/10/big-data/expert-interview-series-erki-kert-big-data-scoring/ [Accessed 28 Mar. 2017].

Jolls, C. (2001). Antidiscrimination and accommodation. Harvard Law Review, 115(2), 642-699.

Kabramson, (2015). Papers of H. Bentley Hahn: The Man Who Invented the 5-Digit ZIP Code. [Online] Available at: http://archiveblog.jfklibrary.org/2015/05/papers-of-h-bentley-hahn-the-man-who-invented-the-zip-code/ [Accessed 12 Jul. 2018].

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805.

Kshetri, N. (2014). Big data′ s impact on privacy, security and consumer welfare. Telecommunications Policy, 38(11), 1134-1145.

Lawson, J. C. (1995). Knowing the score. US Banker, 105, 11-61.

Lenddo, (2016). About Us. [Online] Available at: https://lenddo.com/about.html [Accessed 29 Jun. 2018].

Lenddo, (2016). Lenddo Scoring Factsheet. [Online] Available at: https://www.lenddo.com/pdfs/Lenddo-Scoring-Factsheet-201608.pdf [Accessed 28 Mar. 2017].

Lending Club, (n.d.). How does an online credit marketplace work? [Online] Available at: https://www.lendingclub.com/public/how-peer-lending-works.action [Accessed 9 Jul. 2018].

Lundberg, S. J., & Startz, R. (1983). Private discrimination and social intervention in competitive labor market. The American Economic Review, 73(3), 340-347.

Meadow, W., & Sunstein, C. R. (2001). Statistics, not experts. Duke Law Journal, 51(2), 629-646.

Mester, L. J. (1997). What's the point of credit scoring?. Business review, 3(Sep/Oct), 3-16.

Michy, A. (2015). How to perform a Logistic Regression in R. [Online] Available at: https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/ [Accessed 14 Jul. 2018].

Munnell, A. H., Tootell, G. M., Browne, L. E., & McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. The American Economic Review, 25-53.

O'Neil, C. (2016). Collateral Damage. In Weapons of math destruction: how big data increases inequality and threatens democracy (pp.141-160). Penguin Random House LLC, New York.

Pager, D., & Shepherd, H. (2008). The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annu. Rev. Sociol, 34, 181-209.

Patoka, J. (2018). Can You Make Money With Peer-to-Peer Lending? [Online] Available at: http://thefinancegenie.com/personalfinance/investing/can-make-money-peer-peer-lending [Accessed 9 Jul. 2018].

Phelps, E. S. (1972). The statistical theory of racism and sexism. The american economic review, 62(4), 659-661.

Pope, D. G., & Sydnor, J. R. (2011). What's in a Picture? Evidence of Discrimination from Prosper. com. Journal of Human resources, 46(1), 53-92.

Prime Meridian Income Fund, (2015). Going Beyond FICO: How P2P Lending Platforms Use Big Data to Determine Risk. [Online] Available at: http://www.pmifunds.com/going-beyond-fico-p2p-lending-platforms-use-big-data-determine-risk/ [Accessed 9 Jul. 2018].

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October). Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

Schafer, R., & Ladd, H. F. (1981). Discrimination in mortgage lending. MIT Press (MA).

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.

STHDA, (n.d.). Cox Proportional-Hazards Model. [Online] Available at: http://www.sthda.com/english/wiki/cox-proportional-hazards-model [Accessed 17 Jul. 2018].

Tremoglie, M. P. (2013). Mobile Banking Is More Common Among Blacks and Hispanics. [Online] Available at: h www.mainstreet.com/article/smart-spending/technology/mobile-banking-morecommon-among-blacks-and-hispanics?page=1 [Accessed 1 Jul. 2018].

Turner, M. A., Ross, S. L., Galster, G. C., & Yinger, J. (2002). Discrimination in metropolitan housing markets: National results from Phase I HDS 2000. Washington, DC: US Department of Housing and Urban Development.

US Census Bureau, (2016). American Community Survey. [Online] Available at: https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2016// [Accessed 12 Jul. 2018].

Yoshino, K. (2011). The new equal protection. Harvard Law Review, 124(3), 747-803.

Yu, P., McLaughlin, J., & Levy, M. (2014). Big Data: A Big Disappointment for Scoring Consumer Credit Risk. NCLC, National Consumer Law Center, Boston, MA, 14.

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values, 41(1), 118-132.

Zarsky, T. Z. (2014). Understanding discrimination in the scored society. Wash. L. Rev., 89, 1375.

R Script

Sun Aug 05 17:22:24 2018

```
######################################################################
#Summary Statistics#####################################################
######################################################################

setwd("C:/Users/orphe/OneDrive - Université Libre de Bruxelles/MA2/Thesis/DATABASE/DATA/2016/Accepted")
loans_2016=read.csv("LoanStats_2016.csv",stringsAsFactors=FALSE,header=T,skip=1,sep=";")

#Number of funded loans
dim(loans_2016)

## [1] 434371    148

#Number of funded loans we keep for our analysis
loans_2016=subset(loans_2016,african_american>0)
dim(loans_2016)

## [1] 113126    148

#Average funded loan size
summary(loans_2016$funded_amnt)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    7500   12000   14446   20000   40000

#Term lenght
table(loans_2016$term)

##
##  36 months  60 months
##      86698      26428

#Average interest rate
summary(loans_2016$int_rate)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0500  0.0900  0.1200  0.1287  0.1600  0.3100

#Median annual income
summary(loans_2016$annual_inc)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   47000   65000   79185   92000 9550000

#Median DTI
summary(loans_2016$dti)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   11.94   17.89   18.50   24.50  999.00      16

#Owns a home
table(loans_2016$home_ownership)

##
##      ANY MORTGAGE      OWN     RENT
##       24    47052    11670    54380

#Loans status
table(loans_2016$loan_status)

##
##        Charged Off            Current            Default
##              13239              59031                  3
##         Fully Paid    In Grace Period  Late (16-30 days)
##              37566               1164                390
## Late (31-120 days)
##               1733

#Loans grades
table(loans_2016$grade)

##
##     A     B     C     D     E     F     G
## 18655 35392 34164 15293  6576  2396   650

#########################################################################
#Acceptance Rate#########################################################
#########################################################################

#Consolidation of accepted and rejected application into one file
setwd("C:/Users/orphe/OneDrive - Université Libre de Bruxelles/MA2/Thesis/DAT
ABASE/DATA/2016/Rejected")
loans_R_2016Q1=read.csv("RejectStats_2016Q1.csv",stringsAsFactors=FALSE,heade
r=T,skip=1,sep=";")
loans_R_2016Q2=read.csv("RejectStats_2016Q2.csv",stringsAsFactors=FALSE,heade
r=T,skip=1,sep=";")
loans_R_2016Q3=read.csv("RejectStats_2016Q3.csv",stringsAsFactors=FALSE,heade
r=T,skip=1,sep=";")
loans_R_2016Q4=read.csv("RejectStats_2016Q4.csv",stringsAsFactors=FALSE,heade
r=T,skip=1,sep=";")
loans_A_2016=read.csv("Loans_2016_accepted.csv",stringsAsFactors=FALSE,header
=T,skip=1,sep=";")
loans_A_2016$Debt.To.Income.Ratio=loans_A_2016$Debt.To.Income.Ratio/100
loans_F_2016=rbind(loans_R_2016Q1,loans_R_2016Q2,loans_R_2016Q3,loans_R_2016Q
4,loans_A_2016)

#########################################################################
#Summary Statistics#####################################################
```

```r
#Number of applications
dim(loans_F_2016)
```

```
## [1] 4576653       9
```

```r
#Number of applications we keep for our analysis
loans_F_2016=subset(loans_F_2016,african_american>0)
dim(loans_F_2016)
```

```
## [1] 1224091       9
```

```r
#Acceptance rates
#Policy.Code takes the value 1 when the loan is accepted and 0 otherwise
black_F=subset(loans_F_2016,dummy==1)
not_black_F=subset(loans_F_2016,dummy==0)
sum(black_F$Policy.Code)/length(black_F[,1])
```

```
## [1] 0.07558159
```

```r
sum(not_black_F$Policy.Code)/length(not_black_F[,1])
```

```
## [1] 0.09470417
```

```r
#########################################################################
#Logistic Regressions#####################################################

#We delete rows with na
loans_F_2016=loans_F_2016[complete.cases(loans_F_2016$Debt.To.Income.Ratio),]

#Regression with only control variables
l_reg_c=glm(loans_F_2016$Policy.Code~loans_F_2016$loan_amnt+loans_F_2016$Debt
.To.Income.Ratio+loans_F_2016$emp_year, family = binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(l_reg_c)
```

```
##
## Call:
## glm(formula = loans_F_2016$Policy.Code ~ loans_F_2016$loan_amnt +
##     loans_F_2016$Debt.To.Income.Ratio + loans_F_2016$emp_year,
##     family = binomial())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5664  -0.2680  -0.2499  -0.2366   4.0715
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -3.443e+00  7.227e-03 -476.49   <2e-16
## loans_F_2016$loan_amnt             6.731e-06  1.765e-07   38.14   <2e-16
## loans_F_2016$Debt.To.Income.Ratio -5.085e-01  1.792e-02  -28.37   <2e-16
```

```
## loans_F_2016$emp_year              4.017e-01  1.031e-03  389.65   <2e-16
##
## (Intercept)                       ***
## loans_F_2016$loan_amnt            ***
## loans_F_2016$Debt.To.Income.Ratio ***
## loans_F_2016$emp_year             ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 754191  on 1224074  degrees of freedom
## Residual deviance: 567006  on 1224071  degrees of freedom
## AIC: 567014
##
## Number of Fisher Scoring iterations: 11

library(DescTools)

## Warning: package 'DescTools' was built under R version 3.4.4

PseudoR2(l_reg_c)

##  McFadden
## 0.2481942

#Regression with dummy variable
l_reg_d=glm(loans_F_2016$Policy.Code~loans_F_2016$dummy+loans_F_2016$loan_amn
t+loans_F_2016$Debt.To.Income.Ratio+loans_F_2016$emp_year, family = binomial(
link = 'logit'))

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(l_reg_d)

##
## Call:
## glm(formula = loans_F_2016$Policy.Code ~ loans_F_2016$dummy +
##     loans_F_2016$loan_amnt + loans_F_2016$Debt.To.Income.Ratio +
##     loans_F_2016$emp_year, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5763  -0.2699  -0.2513  -0.2334   4.0649
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -3.418e+00  7.328e-03 -466.35   <2e-16
## loans_F_2016$dummy                -2.277e-01  1.173e-02  -19.40   <2e-16
## loans_F_2016$loan_amnt             6.665e-06  1.770e-07   37.65   <2e-16
## loans_F_2016$Debt.To.Income.Ratio -5.082e-01  1.793e-02  -28.35   <2e-16
## loans_F_2016$emp_year              4.017e-01  1.031e-03  389.44   <2e-16
```

```
##
## (Intercept)                         ***
## loans_F_2016$dummy                  ***
## loans_F_2016$loan_amnt              ***
## loans_F_2016$Debt.To.Income.Ratio ***
## loans_F_2016$emp_year               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 754191  on 1224074  degrees of freedom
## Residual deviance: 566615  on 1224070  degrees of freedom
## AIC: 566625
##
## Number of Fisher Scoring iterations: 11
```

```
library(DescTools)
PseudoR2(l_reg_d)
```

```
##  McFadden
## 0.2487117
```

```
#Regression with relationships between dummy variable and other variables
l_reg_cd=glm(loans_F_2016$Policy.Code~loans_F_2016$dummy+loans_F_2016$loan_am
nt+loans_F_2016$Debt.To.Income.Ratio+loans_F_2016$emp_year
               +loans_F_2016$dummy*loans_F_2016$loan_amnt+loans_F_2016$dummy*
loans_F_2016$Debt.To.Income.Ratio+loans_F_2016$dummy*loans_F_2016$emp_year, f
amily = binomial(link = 'logit'))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(l_reg_cd)
```

```
##
## Call:
## glm(formula = loans_F_2016$Policy.Code ~ loans_F_2016$dummy +
##     loans_F_2016$loan_amnt + loans_F_2016$Debt.To.Income.Ratio +
##     loans_F_2016$emp_year + loans_F_2016$dummy * loans_F_2016$loan_amnt +
##     loans_F_2016$dummy * loans_F_2016$Debt.To.Income.Ratio +
##     loans_F_2016$dummy * loans_F_2016$emp_year, family = binomial(link = "
logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5669  -0.2707  -0.2526  -0.2330   4.1051
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                      -3.401e+00  7.626e-03
## loans_F_2016$dummy                               -3.911e-01  2.403e-02
```

```
## loans_F_2016$loan_amnt                                   6.406e-06  1.873e-07
## loans_F_2016$Debt.To.Income.Ratio                       -5.254e-01  1.909e-02
## loans_F_2016$emp_year                                    3.993e-01  1.087e-03
## loans_F_2016$dummy:loans_F_2016$loan_amnt                2.636e-06  5.756e-07
## loans_F_2016$dummy:loans_F_2016$Debt.To.Income.Ratio     1.534e-01  5.502e-02
## loans_F_2016$dummy:loans_F_2016$emp_year                 2.294e-02  3.443e-03
##                                                          z value Pr(>|z|)
## (Intercept)                                             -445.947  < 2e-16 ***
## loans_F_2016$dummy                                       -16.275  < 2e-16 ***
## loans_F_2016$loan_amnt                                    34.196  < 2e-16 ***
## loans_F_2016$Debt.To.Income.Ratio                        -27.523  < 2e-16 ***
## loans_F_2016$emp_year                                    367.383  < 2e-16 ***
## loans_F_2016$dummy:loans_F_2016$loan_amnt                  4.580 4.65e-06 ***
## loans_F_2016$dummy:loans_F_2016$Debt.To.Income.Ratio       2.788   0.0053 **
## loans_F_2016$dummy:loans_F_2016$emp_year                   6.664 2.66e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 754191  on 1224074  degrees of freedom
## Residual deviance: 566548  on 1224067  degrees of freedom
## AIC: 566564
##
## Number of Fisher Scoring iterations: 11
```

```
library(DescTools)
PseudoR2(l_reg_cd)
```

```
##  McFadden
## 0.2488011
```

```
#########################################################################
#Interest Rate#############################################################
#########################################################################

#########################################################################
#Summary Statistics#######################################################

#Interest rates
black=subset(loans_2016,dummy==1)
mean(black$int_rate)
```

```
## [1] 0.1311808
```

```
not_black=subset(loans_2016,dummy==0)
mean(not_black$int_rate)
```

```
## [1] 0.1284463
```

```
#########################################################################
#OLS Regressions##########################################################
```

```r
setwd("C:/Users/orphe/OneDrive - Université Libre de Bruxelles/MA2/Thesis/DATABASE/DATA/2016/Accepted")
loans_2016_IR=read.csv("LoanStats_2016_Int_Rate.csv",stringsAsFactors=FALSE,header=T,skip=1,sep=";")
loans_2016_IR=subset(loans_2016_IR,african_american>0)

#We delete rows with na
loans_2016_IR=loans_2016_IR[complete.cases(loans_2016_IR$inq_last_6mths),]
loans_2016_IR=loans_2016_IR[complete.cases(loans_2016_IR$revol_util),]

#Regression with only control variables
OLS_c=lm(loans_2016_IR$int_rate~loans_2016_IR$loan_amnt+loans_2016_IR$term
       +loans_2016_IR$emp_length+loans_2016_IR$home_ownership
       +loans_2016_IR$annual_inc+loans_2016_IR$dti+loans_2016_IR$delinq_2yrs
       +loans_2016_IR$earliest_cr_line+loans_2016_IR$inq_last_6mths
       +loans_2016_IR$open_acc+loans_2016_IR$pub_rec+loans_2016_IR$revol_bal
       +loans_2016_IR$revol_util+loans_2016_IR$total_acc)
summary(OLS_c)
```

```
##
## Call:
## lm(formula = loans_2016_IR$int_rate ~ loans_2016_IR$loan_amnt +
##     loans_2016_IR$term + loans_2016_IR$emp_length + loans_2016_IR$home_ownership +
##     loans_2016_IR$annual_inc + loans_2016_IR$dti + loans_2016_IR$delinq_2yrs +
##     loans_2016_IR$earliest_cr_line + loans_2016_IR$inq_last_6mths +
##     loans_2016_IR$open_acc + loans_2016_IR$pub_rec + loans_2016_IR$revol_bal +
##     loans_2016_IR$revol_util + loans_2016_IR$total_acc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89609 -0.02902 -0.00623  0.02289  0.25629
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    9.781e-02  5.396e-04 181.254  < 2e-16 ***
## loans_2016_IR$loan_amnt        4.281e-07  1.628e-08  26.290  < 2e-16 ***
## loans_2016_IR$term             4.241e-02  3.150e-04 134.647  < 2e-16 ***
## loans_2016_IR$emp_length      -2.143e-04  3.248e-05  -6.599 4.17e-11 ***
## loans_2016_IR$home_ownership   4.992e-03  4.071e-04  12.261  < 2e-16 ***
## loans_2016_IR$annual_inc      -3.370e-08  1.638e-09 -20.574  < 2e-16 ***
## loans_2016_IR$dti              9.676e-02  1.397e-03  69.263  < 2e-16 ***
## loans_2016_IR$delinq_2yrs      2.883e-03  1.314e-04  21.934  < 2e-16 ***
## loans_2016_IR$earliest_cr_line -8.509e-04  1.722e-05 -49.425  < 2e-16 ***
## loans_2016_IR$inq_last_6mths   1.306e-02  1.472e-04  88.729  < 2e-16 ***
## loans_2016_IR$open_acc         2.377e-04  3.193e-05   7.445 9.74e-14 ***
## loans_2016_IR$pub_rec          5.405e-03  1.950e-04  27.710  < 2e-16 ***
```

```
## loans_2016_IR$revol_bal        -1.829e-07  5.904e-09 -30.983  < 2e-16 ***
## loans_2016_IR$revol_util        3.621e-02  5.498e-04  65.868  < 2e-16 ***
## loans_2016_IR$total_acc        -4.914e-04  1.556e-05 -31.580  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0415 on 113039 degrees of freedom
## Multiple R-squared:  0.3021, Adjusted R-squared:  0.302
## F-statistic:  3495 on 14 and 113039 DF,  p-value: < 2.2e-16

#Regression with dummy variable
OLS_d=lm(loans_2016_IR$int_rate~loans_2016_IR$loan_amnt+loans_2016_IR$term
         +loans_2016_IR$emp_length+loans_2016_IR$home_ownership
         +loans_2016_IR$annual_inc+loans_2016_IR$dti+loans_2016_IR$delinq_2yr
s
         +loans_2016_IR$earliest_cr_line+loans_2016_IR$inq_last_6mths
         +loans_2016_IR$open_acc+loans_2016_IR$pub_rec+loans_2016_IR$revol_ba
l
         +loans_2016_IR$revol_util+loans_2016_IR$total_acc+loans_2016_IR$dumm
y)
summary(OLS_d)

##
## Call:
## lm(formula = loans_2016_IR$int_rate ~ loans_2016_IR$loan_amnt +
##     loans_2016_IR$term + loans_2016_IR$emp_length + loans_2016_IR$home_own
ership +
##     loans_2016_IR$annual_inc + loans_2016_IR$dti + loans_2016_IR$delinq_2y
rs +
##     loans_2016_IR$earliest_cr_line + loans_2016_IR$inq_last_6mths +
##     loans_2016_IR$open_acc + loans_2016_IR$pub_rec + loans_2016_IR$revol_b
al +
##     loans_2016_IR$revol_util + loans_2016_IR$total_acc + loans_2016_IR$dum
my)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89536 -0.02903 -0.00622  0.02290  0.25598
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    9.768e-02  5.404e-04 180.773  < 2e-16 ***
## loans_2016_IR$loan_amnt         4.292e-07  1.628e-08  26.356  < 2e-16 ***
## loans_2016_IR$term              4.240e-02  3.150e-04 134.603  < 2e-16 ***
## loans_2016_IR$emp_length       -2.157e-04  3.248e-05  -6.642 3.11e-11 ***
## loans_2016_IR$home_ownership    4.922e-03  4.074e-04  12.081  < 2e-16 ***
## loans_2016_IR$annual_inc       -3.363e-08  1.638e-09 -20.533  < 2e-16 ***
## loans_2016_IR$dti               9.670e-02  1.397e-03  69.220  < 2e-16 ***
## loans_2016_IR$delinq_2yrs       2.875e-03  1.314e-04  21.871  < 2e-16 ***
## loans_2016_IR$earliest_cr_line -8.517e-04  1.721e-05 -49.475  < 2e-16 ***
```

```
## loans_2016_IR$inq_last_6mths    1.307e-02  1.472e-04  88.758  < 2e-16 ***
## loans_2016_IR$open_acc          2.370e-04  3.193e-05   7.423 1.16e-13 ***
## loans_2016_IR$pub_rec           5.379e-03  1.951e-04  27.570  < 2e-16 ***
## loans_2016_IR$revol_bal        -1.828e-07  5.903e-09 -30.963  < 2e-16 ***
## loans_2016_IR$revol_util        3.621e-02  5.498e-04  65.864  < 2e-16 ***
## loans_2016_IR$total_acc        -4.919e-04  1.556e-05 -31.615  < 2e-16 ***
## loans_2016_IR$dummy             1.854e-03  4.162e-04   4.455 8.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04149 on 113038 degrees of freedom
## Multiple R-squared:  0.3022, Adjusted R-squared:  0.3021
## F-statistic:  3264 on 15 and 113038 DF,  p-value: < 2.2e-16

#Regression with relationships between dummy variable and other variables
OLS_cd=lm(loans_2016_IR$int_rate~loans_2016_IR$loan_amnt+loans_2016_IR$term
          +loans_2016_IR$emp_length+loans_2016_IR$home_ownership
          +loans_2016_IR$annual_inc+loans_2016_IR$dti+loans_2016_IR$delinq_2
yrs
          +loans_2016_IR$earliest_cr_line+loans_2016_IR$inq_last_6mths
          +loans_2016_IR$open_acc+loans_2016_IR$pub_rec+loans_2016_IR$revol_
bal
          +loans_2016_IR$revol_util+loans_2016_IR$total_acc+loans_2016_IR$du
mmy
          +loans_2016_IR$loan_amnt*loans_2016_IR$dummy+loans_2016_IR$term*lo
ans_2016_IR$dummy
          +loans_2016_IR$emp_length*loans_2016_IR$dummy+loans_2016_IR$home_o
wnership*loans_2016_IR$dummy
          +loans_2016_IR$annual_inc*loans_2016_IR$dummy+loans_2016_IR$dti*lo
ans_2016_IR$dummy+loans_2016_IR$delinq_2yrs*loans_2016_IR$dummy
          +loans_2016_IR$earliest_cr_line*loans_2016_IR$dummy+loans_2016_IR$
inq_last_6mths*loans_2016_IR$dummy
          +loans_2016_IR$open_acc*loans_2016_IR$dummy+loans_2016_IR$pub_rec*
loans_2016_IR$dummy+loans_2016_IR$revol_bal*loans_2016_IR$dummy
          +loans_2016_IR$revol_util*loans_2016_IR$dummy+loans_2016_IR$total_
acc*loans_2016_IR$dummy)
summary(OLS_cd)

##
## Call:
## lm(formula = loans_2016_IR$int_rate ~ loans_2016_IR$loan_amnt +
##     loans_2016_IR$term + loans_2016_IR$emp_length + loans_2016_IR$home_own
ership +
##     loans_2016_IR$annual_inc + loans_2016_IR$dti + loans_2016_IR$delinq_2y
rs +
##     loans_2016_IR$earliest_cr_line + loans_2016_IR$inq_last_6mths +
##     loans_2016_IR$open_acc + loans_2016_IR$pub_rec + loans_2016_IR$revol_b
al +
##     loans_2016_IR$revol_util + loans_2016_IR$total_acc + loans_2016_IR$dum
my +
```

```
##     loans_2016_IR$loan_amnt * loans_2016_IR$dummy + loans_2016_IR$term *
##     loans_2016_IR$dummy + loans_2016_IR$emp_length * loans_2016_IR$dummy +
##     loans_2016_IR$home_ownership * loans_2016_IR$dummy + loans_2016_IR$ann
ual_inc *
##     loans_2016_IR$dummy + loans_2016_IR$dti * loans_2016_IR$dummy +
##     loans_2016_IR$delinq_2yrs * loans_2016_IR$dummy + loans_2016_IR$earlie
st_cr_line *
##     loans_2016_IR$dummy + loans_2016_IR$inq_last_6mths * loans_2016_IR$dum
my +
##     loans_2016_IR$open_acc * loans_2016_IR$dummy + loans_2016_IR$pub_rec *
##     loans_2016_IR$dummy + loans_2016_IR$revol_bal * loans_2016_IR$dummy +
##     loans_2016_IR$revol_util * loans_2016_IR$dummy + loans_2016_IR$total_a
cc *
##     loans_2016_IR$dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87902 -0.02902 -0.00622  0.02289  0.24695
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                       9.807e-02  5.677e-04
## loans_2016_IR$loan_amnt                           4.202e-07  1.709e-08
## loans_2016_IR$term                                4.236e-02  3.318e-04
## loans_2016_IR$emp_length                         -2.108e-04  3.424e-05
## loans_2016_IR$home_ownership                      4.593e-03  4.361e-04
## loans_2016_IR$annual_inc                         -3.241e-08  1.670e-09
## loans_2016_IR$dti                                 9.505e-02  1.455e-03
## loans_2016_IR$delinq_2yrs                         2.864e-03  1.393e-04
## loans_2016_IR$earliest_cr_line                   -8.506e-04  1.821e-05
## loans_2016_IR$inq_last_6mths                      1.302e-02  1.549e-04
## loans_2016_IR$open_acc                            2.396e-04  3.382e-05
## loans_2016_IR$pub_rec                             5.396e-03  2.089e-04
## loans_2016_IR$revol_bal                          -1.819e-07  6.252e-09
## loans_2016_IR$revol_util                          3.639e-02  5.786e-04
## loans_2016_IR$total_acc                          -4.988e-04  1.645e-05
## loans_2016_IR$dummy                              -1.511e-03  1.852e-03
## loans_2016_IR$loan_amnt:loans_2016_IR$dummy       1.303e-07  5.705e-08
## loans_2016_IR$term:loans_2016_IR$dummy            2.257e-04  1.056e-03
## loans_2016_IR$emp_length:loans_2016_IR$dummy     -2.907e-05  1.080e-04
## loans_2016_IR$home_ownership:loans_2016_IR$dummy  2.526e-03  1.224e-03
## loans_2016_IR$annual_inc:loans_2016_IR$dummy     -3.131e-08  8.558e-09
## loans_2016_IR$dti:loans_2016_IR$dummy             1.756e-02  5.241e-03
## loans_2016_IR$delinq_2yrs:loans_2016_IR$dummy     1.593e-04  4.198e-04
## loans_2016_IR$earliest_cr_line:loans_2016_IR$dummy -1.522e-05  5.594e-05
## loans_2016_IR$inq_last_6mths:loans_2016_IR$dummy  4.683e-04  4.971e-04
## loans_2016_IR$open_acc:loans_2016_IR$dummy       -4.334e-05  1.027e-04
## loans_2016_IR$pub_rec:loans_2016_IR$dummy        -7.105e-05  5.854e-04
## loans_2016_IR$revol_bal:loans_2016_IR$dummy       9.190e-09  1.957e-08
## loans_2016_IR$revol_util:loans_2016_IR$dummy     -2.221e-03  1.857e-03
```

```
## loans_2016_IR$total_acc:loans_2016_IR$dummy         7.347e-05  5.077e-05
##                                                    t value Pr(>|t|)
## (Intercept)                                        172.758  < 2e-16 ***
## loans_2016_IR$loan_amnt                             24.589  < 2e-16 ***
## loans_2016_IR$term                                 127.685  < 2e-16 ***
## loans_2016_IR$emp_length                            -6.156  7.5e-10 ***
## loans_2016_IR$home_ownership                        10.532  < 2e-16 ***
## loans_2016_IR$annual_inc                           -19.408  < 2e-16 ***
## loans_2016_IR$dti                                   65.310  < 2e-16 ***
## loans_2016_IR$delinq_2yrs                           20.556  < 2e-16 ***
## loans_2016_IR$earliest_cr_line                     -46.724  < 2e-16 ***
## loans_2016_IR$inq_last_6mths                        84.063  < 2e-16 ***
## loans_2016_IR$open_acc                               7.085  1.4e-12 ***
## loans_2016_IR$pub_rec                               25.836  < 2e-16 ***
## loans_2016_IR$revol_bal                            -29.088  < 2e-16 ***
## loans_2016_IR$revol_util                            62.895  < 2e-16 ***
## loans_2016_IR$total_acc                            -30.324  < 2e-16 ***
## loans_2016_IR$dummy                                 -0.816 0.414713
## loans_2016_IR$loan_amnt:loans_2016_IR$dummy          2.284 0.022395 *
## loans_2016_IR$term:loans_2016_IR$dummy               0.214 0.830827
## loans_2016_IR$emp_length:loans_2016_IR$dummy        -0.269 0.787781
## loans_2016_IR$home_ownership:loans_2016_IR$dummy     2.064 0.039019 *
## loans_2016_IR$annual_inc:loans_2016_IR$dummy        -3.659 0.000253 ***
## loans_2016_IR$dti:loans_2016_IR$dummy                3.349 0.000810 ***
## loans_2016_IR$delinq_2yrs:loans_2016_IR$dummy        0.379 0.704329
## loans_2016_IR$earliest_cr_line:loans_2016_IR$dummy  -0.272 0.785586
## loans_2016_IR$inq_last_6mths:loans_2016_IR$dummy     0.942 0.346152
## loans_2016_IR$open_acc:loans_2016_IR$dummy          -0.422 0.672967
## loans_2016_IR$pub_rec:loans_2016_IR$dummy           -0.121 0.903394
## loans_2016_IR$revol_bal:loans_2016_IR$dummy          0.470 0.638690
## loans_2016_IR$revol_util:loans_2016_IR$dummy        -1.196 0.231569
## loans_2016_IR$total_acc:loans_2016_IR$dummy          1.447 0.147868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04149 on 113024 degrees of freedom
## Multiple R-squared:  0.3025, Adjusted R-squared:  0.3024
## F-statistic:  1691 on 29 and 113024 DF,  p-value: < 2.2e-16

######################################################################
#Default Rate#########################################################
######################################################################

#install.packages(c("survival", "survminer"))
setwd("C:/Users/orphe/OneDrive - Université Libre de Bruxelles/MA2/Thesis/DAT
ABASE/DATA/2016/Accepted")
loans_2016_D=read.csv("LoanStats_2016_Default.csv",stringsAsFactors=FALSE,hea
der=T,skip=1,sep=";",na.strings=c(""))
loans_2016_D=subset(loans_2016_D,african_american>0)
dim(loans_2016_D)
```

```
## [1] 113126    150

#Delete rows with na
loans_2016_D=loans_2016_D[complete.cases(loans_2016_D$inq_last_6mths),]
loans_2016_D=loans_2016_D[complete.cases(loans_2016_D$revol_util),]

#Transform time in numeric and delete 153 na
time_d=as.numeric(loans_2016_D$time)

## Warning: NAs introduits lors de la conversion automatique

loans_2016_D=loans_2016_D[complete.cases(time_d),]

######################################################################
#Summary Statistics#####################################################

#Suppression of In Grace Period
loans_2016_D=loans_2016_D[loans_2016_D$loan_status!="In Grace Period",]

#Default rates
black=subset(loans_2016_D,dummy==1)
not_black=subset(loans_2016_D,dummy==0)
sum(black$Default==1)/length(black[,1])

## [1] 0.1229516

sum(not_black$Default==1)/length(not_black[,1])

## [1] 0.1164422

######################################################################
#Cox Proportional-Hazards Regressions#################################

#Suppression of Fully Paid
loans_2016_D=loans_2016_D[loans_2016_D$loan_status!="Fully Paid",]
dim(loans_2016_D)

## [1] 74203   150

#Regression with only control variables
library("survival")

## Warning: package 'survival' was built under R version 3.4.4

c_reg_c=coxph(Surv(as.numeric(loans_2016_D$time),loans_2016_D$Default)~loans_
2016_D$loan_amnt+loans_2016_D$term
                    +loans_2016_D$emp_length+loans_2016_D$home_ownership
                    +loans_2016_D$annual_inc+loans_2016_D$dti+loans_2016_D$d
elinq_2yrs
                    +loans_2016_D$earliest_cr_line+loans_2016_D$inq_last_6mt
hs
                    +loans_2016_D$open_acc+loans_2016_D$pub_rec+loans_2016_D
```

```
$revol_bal
                    +loans_2016_D$revol_util+loans_2016_D$total_acc)
summary(c_reg_c)

## Call:
## coxph(formula = Surv(as.numeric(loans_2016_D$time), loans_2016_D$Default)
~
##     loans_2016_D$loan_amnt + loans_2016_D$term + loans_2016_D$emp_length +
##         loans_2016_D$home_ownership + loans_2016_D$annual_inc +
##         loans_2016_D$dti + loans_2016_D$delinq_2yrs + loans_2016_D$earlies
t_cr_line +
##         loans_2016_D$inq_last_6mths + loans_2016_D$open_acc +
##         loans_2016_D$pub_rec + loans_2016_D$revol_bal + loans_2016_D$revol
_util +
##         loans_2016_D$total_acc)
##
##   n= 74203, number of events= 13082
##
##                                       coef  exp(coef)   se(coef)       z
## loans_2016_D$loan_amnt          2.514e-05  1.000e+00  1.269e-06  19.810
## loans_2016_D$term               9.505e-02  1.100e+00  2.086e-02   4.557
## loans_2016_D$emp_length        -1.071e-02  9.893e-01  2.309e-03  -4.638
## loans_2016_D$home_ownership     4.154e-02  1.042e+00  2.886e-02   1.440
## loans_2016_D$annual_inc        -4.508e-06  1.000e+00  2.727e-07 -16.530
## loans_2016_D$dti                3.402e-01  1.405e+00  2.681e-02  12.690
## loans_2016_D$delinq_2yrs        2.475e-02  1.025e+00  8.549e-03   2.895
## loans_2016_D$earliest_cr_line -1.482e-02  9.853e-01  1.308e-03 -11.327
## loans_2016_D$inq_last_6mths     2.690e-01  1.309e+00  8.851e-03  30.389
## loans_2016_D$open_acc           3.408e-04  1.000e+00  2.188e-03   0.156
## loans_2016_D$pub_rec            8.343e-02  1.087e+00  1.165e-02   7.162
## loans_2016_D$revol_bal         -8.839e-06  1.000e+00  8.203e-07 -10.775
## loans_2016_D$revol_util         3.240e-01  1.383e+00  4.113e-02   7.878
## loans_2016_D$total_acc          7.812e-03  1.008e+00  1.077e-03   7.256
##                                    Pr(>|z|)
## loans_2016_D$loan_amnt          < 2e-16 ***
## loans_2016_D$term              5.18e-06 ***
## loans_2016_D$emp_length        3.51e-06 ***
## loans_2016_D$home_ownership     0.14999
## loans_2016_D$annual_inc         < 2e-16 ***
## loans_2016_D$dti                < 2e-16 ***
## loans_2016_D$delinq_2yrs        0.00379 **
## loans_2016_D$earliest_cr_line   < 2e-16 ***
## loans_2016_D$inq_last_6mths     < 2e-16 ***
## loans_2016_D$open_acc           0.87621
## loans_2016_D$pub_rec           7.96e-13 ***
## loans_2016_D$revol_bal          < 2e-16 ***
## loans_2016_D$revol_util        3.33e-15 ***
## loans_2016_D$total_acc         3.99e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                                exp(coef) exp(-coef) lower .95 upper .95
## loans_2016_D$loan_amnt            1.0000     1.0000    1.0000    1.0000
## loans_2016_D$term                 1.0997     0.9093    1.0557    1.1456
## loans_2016_D$emp_length           0.9893     1.0108    0.9849    0.9938
## loans_2016_D$home_ownership       1.0424     0.9593    0.9851    1.1031
## loans_2016_D$annual_inc           1.0000     1.0000    1.0000    1.0000
## loans_2016_D$dti                  1.4053     0.7116    1.3333    1.4811
## loans_2016_D$delinq_2yrs          1.0251     0.9756    1.0080    1.0424
## loans_2016_D$earliest_cr_line     0.9853     1.0149    0.9828    0.9878
## loans_2016_D$inq_last_6mths       1.3086     0.7642    1.2861    1.3315
## loans_2016_D$open_acc             1.0003     0.9997    0.9961    1.0046
## loans_2016_D$pub_rec              1.0870     0.9200    1.0625    1.1121
## loans_2016_D$revol_bal            1.0000     1.0000    1.0000    1.0000
## loans_2016_D$revol_util           1.3827     0.7232    1.2756    1.4988
## loans_2016_D$total_acc            1.0078     0.9922    1.0057    1.0100
##
## Concordance= 0.62  (se = 0.003 )
## Rsquare= 0.029   (max possible= 0.98 )
## Likelihood ratio test= 2211  on 14 df,   p=<2e-16
## Wald test            = 2343  on 14 df,   p=<2e-16
## Score (logrank) test = 2273  on 14 df,   p=<2e-16
```

```
#Regression with dummy variable
library("survival")
c_reg_d=coxph(Surv(as.numeric(loans_2016_D$time),loans_2016_D$Default)~loans_
2016_D$loan_amnt+loans_2016_D$term
                      +loans_2016_D$emp_length+loans_2016_D$home_ownership
                      +loans_2016_D$annual_inc+loans_2016_D$dti+loans_2016_D$d
elinq_2yrs
                      +loans_2016_D$earliest_cr_line+loans_2016_D$inq_last_6mt
hs
                      +loans_2016_D$open_acc+loans_2016_D$pub_rec+loans_2016_D
$revol_bal
                      +loans_2016_D$revol_util+loans_2016_D$total_acc+loans_20
16_D$dummy)
summary(c_reg_d)
```

```
## Call:
## coxph(formula = Surv(as.numeric(loans_2016_D$time), loans_2016_D$Default)
~
##     loans_2016_D$loan_amnt + loans_2016_D$term + loans_2016_D$emp_length +
##         loans_2016_D$home_ownership + loans_2016_D$annual_inc +
##         loans_2016_D$dti + loans_2016_D$delinq_2yrs + loans_2016_D$earlies
t_cr_line +
##         loans_2016_D$inq_last_6mths + loans_2016_D$open_acc +
##         loans_2016_D$pub_rec + loans_2016_D$revol_bal + loans_2016_D$revol
_util +
##         loans_2016_D$total_acc + loans_2016_D$dummy)
##
```

```
##   n= 74203, number of events= 13082
##
##                                     coef  exp(coef)   se(coef)       z
## loans_2016_D$loan_amnt         2.514e-05  1.000e+00  1.269e-06  19.809
## loans_2016_D$term              9.505e-02  1.100e+00  2.086e-02   4.557
## loans_2016_D$emp_length       -1.071e-02  9.893e-01  2.309e-03  -4.638
## loans_2016_D$home_ownership    4.157e-02  1.042e+00  2.888e-02   1.439
## loans_2016_D$annual_inc       -4.508e-06  1.000e+00  2.727e-07 -16.529
## loans_2016_D$dti               3.402e-01  1.405e+00  2.681e-02  12.690
## loans_2016_D$delinq_2yrs       2.475e-02  1.025e+00  8.549e-03   2.895
## loans_2016_D$earliest_cr_line -1.482e-02  9.853e-01  1.309e-03 -11.325
## loans_2016_D$inq_last_6mths    2.690e-01  1.309e+00  8.851e-03  30.388
## loans_2016_D$open_acc          3.412e-04  1.000e+00  2.188e-03   0.156
## loans_2016_D$pub_rec           8.344e-02  1.087e+00  1.166e-02   7.158
## loans_2016_D$revol_bal        -8.839e-06  1.000e+00  8.204e-07 -10.774
## loans_2016_D$revol_util        3.240e-01  1.383e+00  4.113e-02   7.878
## loans_2016_D$total_acc         7.813e-03  1.008e+00  1.077e-03   7.255
## loans_2016_D$dummy            -6.673e-04  9.993e-01  2.889e-02  -0.023
##                                Pr(>|z|)
## loans_2016_D$loan_amnt          < 2e-16 ***
## loans_2016_D$term              5.18e-06 ***
## loans_2016_D$emp_length        3.53e-06 ***
## loans_2016_D$home_ownership     0.15009
## loans_2016_D$annual_inc         < 2e-16 ***
## loans_2016_D$dti                < 2e-16 ***
## loans_2016_D$delinq_2yrs        0.00379 **
## loans_2016_D$earliest_cr_line   < 2e-16 ***
## loans_2016_D$inq_last_6mths     < 2e-16 ***
## loans_2016_D$open_acc           0.87607
## loans_2016_D$pub_rec           8.18e-13 ***
## loans_2016_D$revol_bal          < 2e-16 ***
## loans_2016_D$revol_util        3.33e-15 ***
## loans_2016_D$total_acc         4.00e-13 ***
## loans_2016_D$dummy              0.98157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef) exp(-coef) lower .95 upper .95
## loans_2016_D$loan_amnt            1.0000     1.0000    1.0000    1.0000
## loans_2016_D$term                 1.0997     0.9093    1.0557    1.1456
## loans_2016_D$emp_length           0.9893     1.0108    0.9849    0.9938
## loans_2016_D$home_ownership       1.0424     0.9593    0.9851    1.1032
## loans_2016_D$annual_inc           1.0000     1.0000    1.0000    1.0000
## loans_2016_D$dti                  1.4053     0.7116    1.3333    1.4811
## loans_2016_D$delinq_2yrs          1.0251     0.9756    1.0080    1.0424
## loans_2016_D$earliest_cr_line     0.9853     1.0149    0.9828    0.9878
## loans_2016_D$inq_last_6mths       1.3086     0.7642    1.2861    1.3315
## loans_2016_D$open_acc             1.0003     0.9997    0.9961    1.0046
## loans_2016_D$pub_rec              1.0870     0.9200    1.0625    1.1121
## loans_2016_D$revol_bal            1.0000     1.0000    1.0000    1.0000
```

```
## loans_2016_D$revol_util          1.3827     0.7232    1.2756    1.4988
## loans_2016_D$total_acc           1.0078     0.9922    1.0057    1.0100
## loans_2016_D$dummy               0.9993     1.0007    0.9443    1.0575
##
## Concordance= 0.62  (se = 0.003 )
## Rsquare= 0.029   (max possible= 0.98 )
## Likelihood ratio test= 2211  on 15 df,   p=<2e-16
## Wald test            = 2343  on 15 df,   p=<2e-16
## Score (logrank) test = 2273  on 15 df,   p=<2e-16
```