Project Progress Report

CSCI 4502 Project

Scott Young Michael Gilroy Cory Morales Michael Min

Abstract:

This study attempts to determine which companies involved in finances in the United States are either unethical or consumer unfriendly, determined largely by the legal complaints that consumers have about said companies. Another goal of this study is to determine which areas in the United States are associated the most with conflicts between companies and consumers, also based on consumer complaints.

The results of this study show that while states in the midwest of the U.S. generally appear to have the lowest rates of conflict between consumers and companies, states on both coastlines have higher rates. This discrepancy suggests some degree of relaxed legislation in said states or the business patterns of multiple controversial companies. The results also indicate that several companies tend to be involved disproportionately positively with certain issues, suggesting a questionable code of ethics for said companies.

Introduction:

For the first question we have for this data, we wish to determine which states have the highest percentage of consumer complaints relative to population. This question is important in that it may help to determine which states are strict in the laws that govern finances and other economic affairs. It is environment of the United States' financial system as well as potentially highlighting specific troublesome outlier states and their associated companies.

For the second question, we wish to deduce which individual companies are the most tenuous when it comes to customer satisfaction. This question is important in that answering it may aid consumers in avoiding problematic companies. Another important implication with this question is that companies that do appear especially notable in our findings must be notable throughout multiple types of analysis in our data mining; if companies do repeatedly appear as exemplary on multiple fronts (such as multiple types of correlation), it may provide consumers an entire grouping of evidence when it comes to a particular company's wrongdoings.

Related Work:

The previous work that has been done on the Consumer Financial Protection Bureau dataset (CFPB dataset) that we used in this project can really be boiled down into two categories: visualization and statistical analysis. Some financial analysts at *Deloitte* put together a publically available document giving some insight as to things to be taken from this particular dataset, such as trends in types of mortgage complaints over time[1]. *Deloitte* includes an abundance of visual aids to support their claims- histograms, line graphs, and stacked bar charts are all included in their analysis of the CFPB dataset. They make claims based on demographics, such as how age and income have direct correlation with the number of mortgage complaints. They also acknowledge that this dataset is comprised of only *reported* complaints, and that a healthy amount of skepticism should be exercised when evaluating the semantics of the data.

The organization uspirgedfund.org also did an analysis on the CFPB dataset, and has similar tone to the Deloitte analysis. Similar to Deloitte, uspirgedfund.org agrees that mortgages are a pivotal part of this financial complaint dataset[2]. They divided these mortgage complaints into two large categories: customers who were unable to pay, and those who had issues paying. This kind of analysis is interesting because it details the percentage of complaints where the company is at fault.

Data set:

Our dataset is the database of consumer complaints as provided by the Consumer Financial Protection Bureau[3]. It contains complaint submissions dating from December 1st, 2011. On the date of download (February 21st, 2017), the dataset had entries up to February 20th, 2017. The dataset has a total of 720,000 entries and 18 attributes.

The first attribute is "Date received", which is merely the date that day, month, and year that the complaint was recorded into the data set. The time scale ranges from December of 2011 to March of 2017, yet we decided that we would exclude the dates in February of 2017 and March of 2017 due to how we obtained the current iteration of the data in early March and decided it may be best to leave out data entries that were too recent for some of our analysis

The second attribute is "Product", a nominal mandatory attribute that contains the product that a complaint is being targeted towards in a company per each entry. There are 12 different types of included products total, yet this attribute was not a great focus of our mining. However, we did some relatively basic analysis to determine a sense of scale of the frequencies of the product types.

The third attribute is "Sub-Product". Similar to "Product", this attribute holds the same 12 types of products, yet this field is completely optional.

The fourth attribute and one of the most important in the mining is “Issue”. This attribute is both nominal and mandatory and contains the type of complaint the consumer of each entry has towards a company. There are 95 issue types total, meaning that for some of our analysis, we had to decide which of these 95 issue types would be the most significant in terms of determining a company’s quality of ethics.

The fifth attribute is “Sub-Issue”. Similar to “Issue”, this attribute holds the same 12 types of products, yet this field is completely optional.

The sixth attribute is “Consumer Complaint Narrative”. This attribute is optional and holds any free verbal dialogue that a consumer may have said about the case. Because this attribute was so diverse in its content, we chose to exclude it from our mining for the sake of concentrating on more concrete attributes.

The seventh attribute is “Company Public Response”. This attribute is nominal and optional, holding the company’s defense or dialogue about the case.

The eight attribute is “Company”, which is also one of the more important attributes in this analysis. This attribute is nominal and mandatory and simply lists which company is being accused in the given entry.

The ninth attribute is “State”, which holds the state that the company branch accused is based in. This attribute is nominal and mandatory.

The tenth attribute is “ZIP”, which is optional and holds the zipcode of the branch of the accused company.

The eleventh attribute is “Tags”, which is optional holds any special or demographic information about the entry’s writer. Some examples include “Older American” or “Servicemember”.

The twelfth attribute is “Consumer consent provided?” which holds whether or not the consumer gave consent about the company’s response. This nominal attribute held five types, but was not used much in our analysis.

The thirteenth attribute is “Submitted via” which is mandatory, nominal, and simply contains how the complaint was sent to be processed (such as fax, web, phone, etc.). This attribute was excluded from our analysis.

The fourteenth attribute is “Date sent to company”, which simply states the date that the complaint was given to the company.

The fifteenth attribute is “Company response to consumer”, which is a nominal and mandatory attribute that states the way in which the company either resolved or reimbursed the consumer. Some of the responses were positive, with monetary relief. Others were more detrimental, and were only closed with no explanation.

The sixteenth attribute is “Timely response?”, which simply is a yes or no question about whether the company responded quickly to the complaint.

The seventeenth attribute is “Consumer disputed?”, which simply contained a yes or no answer of whether the consumer disputed the company’s response.

The eighteenth attribute is “Complaint ID”, which holds the each entry’s unique ID number.

Main Techniques Applied:

Data Cleaning:

We have finished marking entries that lack values in mandatory attributes (issue, state, company, etc.) as “UNKNOWN” in those entries as well as placing a constant value “NULL” in blank cells under optional attributes that are not completely necessary to evaluating an entry. This cleaning made our data more comprehensible; a prime example is in our analysis of correlation for complaint types with company types. Here, entries that lacked values in the uncleaned data would appear as a blank space, which initially was interpreted as a bug in the code. However, with the cleaned data, we were better able to recognize that the “UNKNOWN” entries for these correlations were not an error in the code or parsing.

Data sampling:

Through a significant amount of effort, we have succeeded in creating an effective means of sampling data for various reasons. The prime reason for the sampling method’s existence was to decrease the time needed for testing new programs with the dataset for more technical problems (like bug fixing).

Tools:

We used Python as a basic framework, using the matplotlib, numpy, and pandas libraries to detect patterns in the data. We also used the matplotlib.pyplot library in order to plot some simpler data analysis tasks.

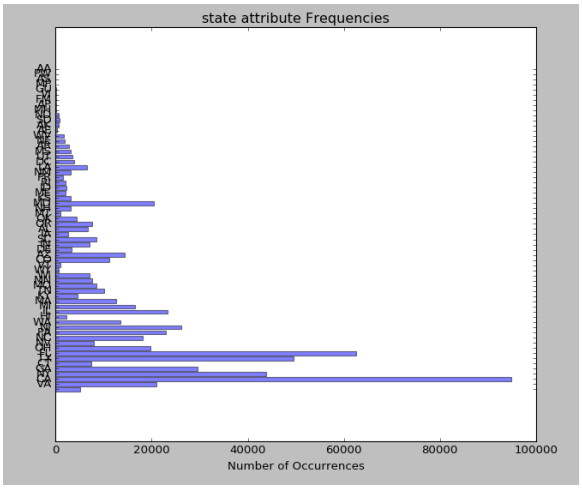
In addition, we also incorporated some use of HTML5 and the Python library urllib2 in order to create mainly the heat maps of the United States.

Finally, for the bar graphs in our results section, we used Jupyter to rapidly form frequencies for every state, product type, or other attribute that we required.

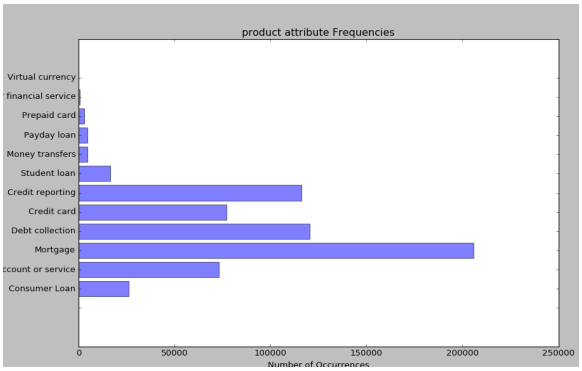
Key Results:

Preliminary graphs simply showed occurrences of various attributes, which prove to be useful when coming

up with tests for more advanced hypotheses. For example, when parsing the data by occurrences by state, we ended up with this graph:

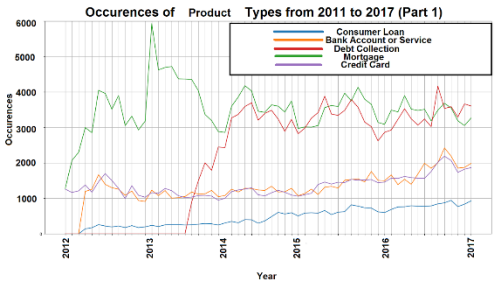


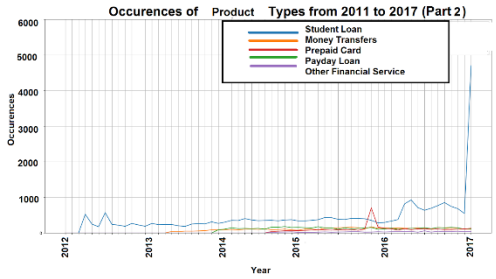
These first visualizations led us to expand upon the location-based findings and make a heatmap of US states and their occurrences. Another big picture finding from initial visualizations was by product type. This graph gave us a high level understanding of what issues we should focus on:



The “mortgage” category has a clear plurality in terms of occurrences among the data, so we were able to shift some questions we asked in a new direction. This graph also was in line with previous research and literature, which gave us some confidence in our parsing of the dataset.

Additionally, we have formulated a pair of scatter plots that measures the number of occurrences of product types over the course of our time frame, which ranged from December of 2011 to January of 2017. The following graphs demonstrate the following growth or decline rates of our product types:

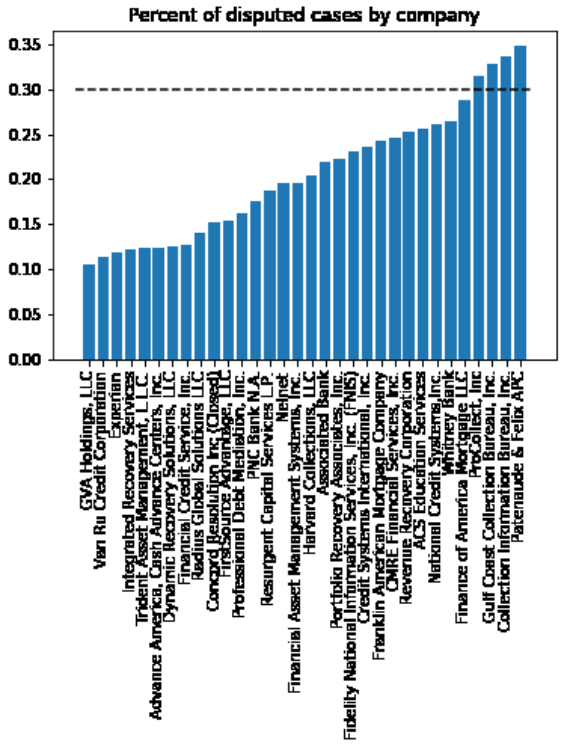




These two scatter plots helped to determine how common product types were in relation to each other as time progressed, as the product types from part one were shown to be dramatically more prevalent from those of part two. Also, these two plots were useful in determining how product type frequencies were affected over time; we believed that these trends could reveal insights on what types of work the companies were involved in.

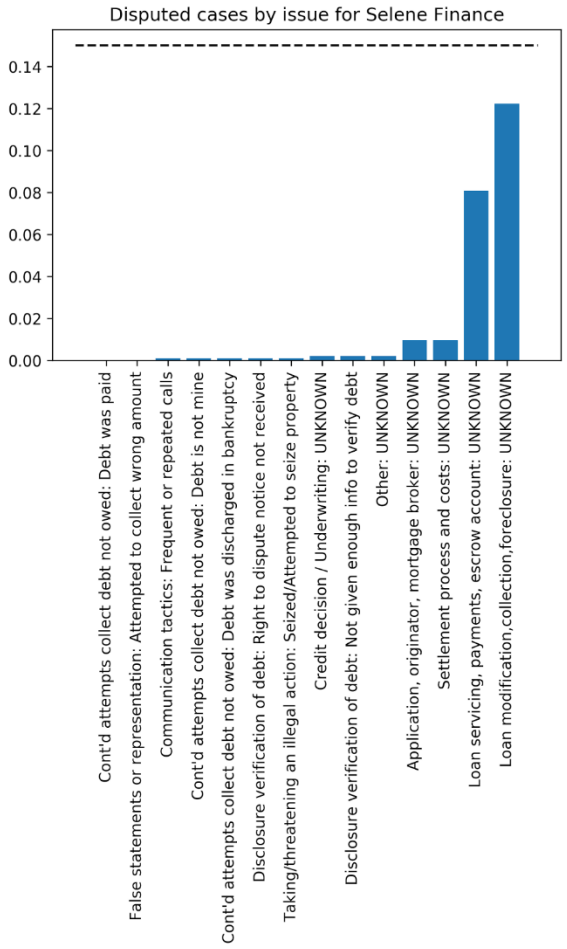
One of the most basic, but perhaps the most informative methods of analysis we used was to look at what fraction of the complaints that a company received were troublesome complaints. The trouble with this was determining which complaints were indicative of malpractice and which ones were not. Due to a lack of thorough knowledge of the relevant field, we used a simple assumption to determine the difference between the two. Each entry contained an attribute for whether or not the consumer disputed the company’s response to the complaints. Although obviously an approximation, we operated under the assumption that complaints where the consumer disputed the claim were the indicative of malpractice, and cases where they didn’t were not indicative of malpractice. Using this assumption we used 2 different methods to pick companies that are potentially committing malpractice.

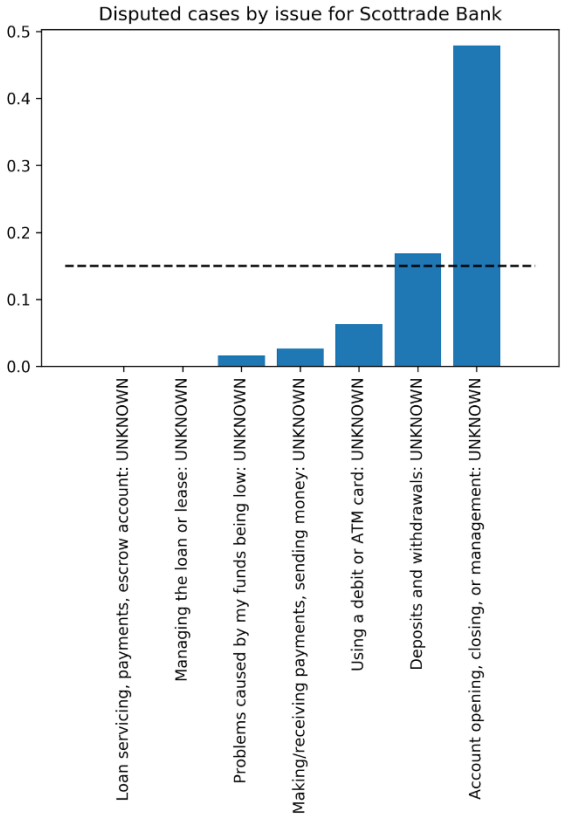
The first was to search for companies with high percentage of the complaints they received ending in dispute. The following is a visualization of a subset of the results from that:



The threshold line indicates which companies had a sufficiently high rate of complaint disputes.

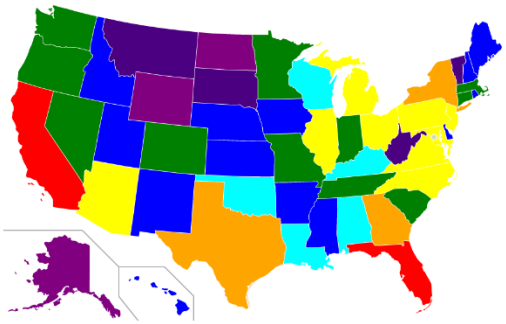
The second method was to search for companies which had a significant number of their disputed complaints be of a specific type. The results from this for a few sample companies can be seen:



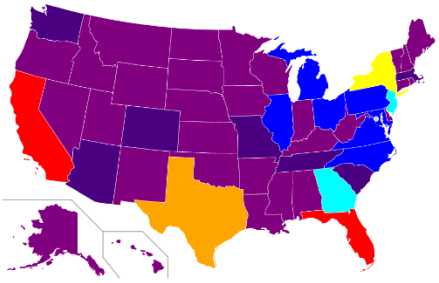


The lines on the graphs again represent the threshold to be included in the list of suspicious companies. Scottrade bank here serves as company that did get detected by this test, while Selene Finance was not detected by this test.

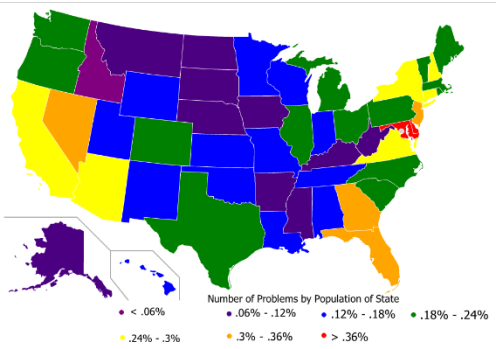
Another completed section of the project involves two separate tasks: implementing additional algorithms to provide a greater range of information on the data and expressing the data and correlations we have currently gathered with a series of graphs. We have acquired a map that color codes ranges of data on the map of the United States. Each color represents a different amount of reported complaints. Purple represents the fewest amount of complaints from 0-1,000. 1,000-2,000 complaints is represented in Indigo, 2,000-4,000 complaints is represented by Blue, 4,000-8,000 is represented by Cyan, 8,000-16,000 is represented by Green, 16,000-32,000 is represented by Yellow, 32,000-64,000 is represented by Orange, and more than 64,000 complaints is represented by Red.



This map helps to visualize how the complaints are distributed around the country. A lot of the complaints come from areas of higher population. Now there is a good distribution of the colors, though the later ones are massively larger than the earlier ones. In order to get a better idea of how the states lie in terms of consumer complaints, we ran the same algorithm with standardized bin sizes. The new ranges were represented by the same colors. Purple represents any state with 0 to 9,150 reported complaints. Indigo represents 9,150 to 18,300 complaints. Blue represents 18,300 to 27,450 complaints. Cyan represents 27,450 to 36,600 complaints. Green represents 36,600 to 45,750 complaints. Yellow represents 45,750 to 54,900 complaints. Orange represents 54,900 to 64,050 complaints, and Red represents any state with more than 64,050 complaints.



With standardized bin sizes of 9,150, we can see that there is a larger disparity of the number of complaints depending on which state one is in. Texas, California and Florida still top the list, and the trend of a higher pocket of complaints from Illinois to the east coast still exists. The only remaining problem is the fact that all of these states vary in terms of population size. In order to get a better picture as to what the map looks, we had to make the map scale with the estimated number of people in the state. Each color represents a different percentage range for the complaints in the state. Purple represents 0% to .06% of the population making a complaint. Indigo represents .06% to .12%, Blue represents .12% to .18%, Green represents .18% to .24%, Yellow represents .24% to .3%, Orange represents .3% to .36%. Any percent higher than .36% is represented in Red.



With the addition of the population based scaling, a more normalized image is seen. California drops several brackets, and Maryland and Delaware skyrocket several brackets. Due to their relatively small population and high number of customer complaints, another question arises. Do certain states have harsher laws in terms of regulating companies, or are companies in certain states being more shady in their practices. After a bit of research, it becomes more clear that the federal government is more responsible for laws that regulate how companies must act. Knowing this, some conclusions can be drawn, but without further investigation into what the customers

were complaining about, there is little basis for the conclusions and possible accusations that would arise.

Some other concrete results at a previous stage included some preliminary information on both lift value and Chi-Square correlations for the data set's nominal attributes.

To clarify, lift correlation values in this context tell us the ratio of the number of times a chosen issue and a company appear together to the total number of times the chosen issue appears at all. This is important to our data mining in that it tells us how commonly a company is responsible for a specific kind issue relative to other companies; a company with a lift correlation value dramatically higher than the average lift for an issue may be a strong sign of that specific company being a prime cause of that issue and therefore, a potential sign of an unethical company. Thus, we used lift correlation widely in this project to determine which companies were primarily responsible for specific issues.

On the other hand we also used Chi-Square correlation alongside lift correlation. Here, Chi-Square is used to determine how likely it was that a particular company and issue appeared as a coincidence in the data. In our code, we determined the Chi-Square value by counting the number of times an issue appeared, an issue did not appear, a company appeared, a company did not appear, and the various combinations of the aforementioned states. Thus, we were able to derive a Chi-Square value for each issue and company pair that we discovered. We also compared the Chi-Square values to the average of all the issue and company pairs for each individual issue in order to determine which pair stood out in the midst of the others' Chi-Square correlations. Similar to the lift correlation, the Chi-Square correlation helped to determine companies that were extreme outliers, or companies that could potentially prove to be exceptionally controversial in our ultimate analysis.

Provided below as an example is a table of lift values between a small sample of individual issues and the company that they are correlated the highest with.

Issue	Company	Lift
Late Fee	FDIC	13.962
Credit Card Protection/Debt Protection	Square One Financial, LLC	92.879
Credit Reporting	Capital One	8.983
Advertising, marketing, or disclosures	NetSpend Corporation, a TSYS Company	423.112
Arbitration	GAMACHE & MYERS, PC	246.555
Cash Advance Fee	Commerce Bank	17.310

In addition, the same is done in the table provided below, but using Chi-Square values instead.

Issue	Company	Chi-Square
Late Fee	Bank of America	6742.708
Credit Card Protection/Debt Protection	Thomas Kerns McKnight, LLP	81608.444
Credit Reporting	Bank of America	6188.154
Advertising, marketing, or disclosures	JPMorgan Chase & Co.	2626.862
Arbitration	GAMACHE & MYERS, PC	9065.975
Cash Advance Fee	Bank of America	6735.540

Thus, with these basic tables, one can determine the types of issues that the displayed companies tend to cause the most.

However, in our more recent stage, we have derived a means of determining the average correlations of all the companies accused for each issue type and the top five most positively correlated companies.

In the case of the previously mentioned six complaint types, we derived the following new tables to see how the data we previously gathered about the the most correlated companies compares to the other companies who were accused of the type of complaint, or for the sake of brevity, the issue types of "Late Fee" and "Credit Card Protection/Debt Protection":

Issue: Late Fee
Average Lift Correlation: 2.738

Company	Lift Correlation
FDIC	13.962
Alliance Data Card Services	13.328
Synchrony Financial	9.120
Green & Cooper Collections LLC	8.726
Lyons, Doughty & Veldhuis, P.C.	7.757

Issue: Credit Card Protection/Debt Protection
Average Lift Correlation: 4.382

Company	Lift Correlation
Square One Financial, LLC	92.879
Novea Portfolio Management, LLC	34.830
APPLE RECOVERY, LLC	18.576
IMC Capital, LLC	15.0614
Boeing Employees' Credit Union	13.592

With this new data visible for life correlation, we can see that all of the companies involved are very strongly correlated with the issues, especially when compared to the average correlation for each of the issues. Thus, it can be deduced this data supports the notion that the aforementioned companies are outstanding with their involvement in the selected issues and are much more correlated with them than other companies are.

In the case of Chi-Square correlations, we did the same for the same issues and produced the following tables:

Issue: Late Fee
Average Chi-Square Correlation: 532.135

Company	Chi-Square Correlation
Bank of America	6094.632
Wells Fargo & Company	3775.741
FDIC	3262.525
Alliance Data Card Services	2935.158
Experian	2886.604

Issue: Credit Card Protection/Debt Protection
Average Chi-Square Correlation: 1431.624

Company	Lift Correlation
Square One Financial, LLC	81608.445
Novea Portfolio Management, LLC	11474.625
Bank of America	6110.549
Wells Fargo & Company	4043.383
APPLE RECOVERY, LLC	3262.525

From these two Chi-Square correlation tables, it is seen again that once again, the most positively correlated companies are dramatically higher than the average correlations. Additionally, it is shown that several companies reappear from the lift correlation graphs, such as Square One Financial, LLC in the issue of "Credit Card Protection/Debt Protection" and FDIC in the issue of "Late Fee". Thus, this Chi-Square analysis further supports the notion that the companies that appear highly in both the lift correlations and Chi-square correlations are even more likely to be unfavorable in reality. On the other hand, companies that appear in the top five list of only one of these types of correlations may require more investigation.

Thus, with the correlation tables displaying a more comprehensive array of data, we were able to establish that the questionable companies were indeed significant in how correlated they were to their issues relative to other companies.

To establish a list of exemplary companies in terms of lift and Chi-Square correlation, instead of using these top five lists for each issue, we instead formed an entire list of all the companies that were the highest correlated with each of the issues (for both of the correlation types as well).

By combining the lists of companies generated by both correlation types and the lists generated by disputes, we found three companies that were prevalent in all of the tests: Amex, Army and Air Force Exchange Service, and FirstBank of Puerto Rico. For each of these companies, provided below are the tables of their best lift-correlated issue type:

Issue: Rewards
Average Lift Correlation: 2.634

Company	Lift Correlation
AMEX	26.155
Barclays PLC	13.999
City National Bank	13.496
First Hawaiian Bank	7.113
Pentagon FCU	6.110

Issue: Collection Debt Dispute
Average Lift Correlation: 2.794

Company	Lift Correlation
Army and Air Force Exchange Service	8.672
Capital One	8.642
Banco Satander Puerto Rico	7.914
Discover	7.911
BancorpSouth Bank	7.344

Issue: Cash Advance
Average Lift Correlation: 4.433

Company	Lift Correlation
FirstBank of Puerto Rico	15.366
First Citizens BancShares, Inc.	12.967
Discover	9.941
Banco Popular North America	8.328
Comerica	8.181

Applications:

Overall, we believe that we can make use of this analysis in many different ways.

Firstly, in the case of the heat map, a consumer can use the data we gathered to determine which states may have more relaxed local laws concerning controversial business strategies. As an example, using the heat map that detects percentages, the states that are yellow, orange, and red (such as Maryland and Delaware) could be considered riskier spots for consumers to do business, such as loans.

Conversely, states that appear in the shades of purple and blue on the same map (which indicate a statewide complaint rate under 0.18%, which is on the lower half of the range of complaint range percentage) could be interpreted by prospective consumers that those states may be safer to do business in for whatever reason, such as stricter laws or fewer malignant companies.

In the case of both correlations (lift and Chi-Square), we believe that we can use this data to determine which companies could be considered unethical when it comes to the specific issues that they are best correlated with. In the previous example with the “Late Fee” issue and the “Credit Card Protection/Debt Protection” issue, the companies that appeared in the top five correlated companies for both correlation types (who were also correlated far above the average correlation values) may be considered unsatisfactory in terms of consumer safety by our standards. Thus, along with our overall list of suspicious companies, the more specialized view of which companies are correlated with each issue could prove useful to a discerning consumer who wishes to know specifically which company appears to be the most common when it comes to a specific issue type related to the consumer’s desired product.

In relation to the grand scope of the data we mined, the three companies that stood out in all of our different means of data analysis (AMEX, Army and Air Force

Exchange Service, and FirstBank of Puerto Rico), may at this point be useful to consumers as companies that should generally be avoided due to the high volume of attributed issues and high correlation values to said issues.

Thus, the ultimate overarching way this mined data can be applied is to determine the general ethical nature of companies that appear in the data set. From there, consumers may be able to avoid dubious companies and decrease a level of confusion while doing their business.

References:

[1]2017. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/se/Documents/financial-services/CFPBConsumerComplaintDatabase091913US_FSI_.pdf. [Accessed: 22- Feb- 2017].

[2]"New Report: Mortgage Problems Rank #1 at CFPB for Consumer Complaints | U.S. PIRG Education Fund", *Uspirgedfund.org*, 2017. [Online]. Available: http://www.uspirgedfund.org/news/usp/new-report-mortgage-problems-rank-1-cfpb-consumer-complaints. [Accessed: 22- Feb- 2017].

[3]"Consumer Complaints | Consumer Financial Protection Bureau", *Consumer Financial Protection Bureau*, 2017. [Online]. Available: https://data.consumerfinance.gov/dataset/Consumer-Complaints/s6ew-h6mp. [Accessed: 21- Feb- 2017].