Data Visualisation On Real Customer Dispute Data

JASKIRAT SINGH BHATIA*, University Of Waterloo

ACM Reference Format:
Jaskirat Singh Bhatia. 2019. Data Visualisation On Real Customer Dispute Data. 1, 1 (December 2019), 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Consumer feedback plays an important role in improving the services which a Company provides. In my project, I worked on a public data-set which had gathered the feed backs of consumers(US Based) for various companies and their products. Using the data, I tried to find some Interesting patterns or visualise what went wrong over the years or what has improved since.

2 DATA SET

It is a Public data-set available here

The data-set contained a total of 1645795 Entries.

2.1 Date Headers

- Date received: The date when The Complaints was recieved
- Product: The Product for which the complaint was filed
- Sub-product: The Sub-Product of the Product
- Issue: Issue faced by the Consumer
- Sub-issue: Sub-Issue of the Issue
- Consumer complaint narrative: What the consumer said in the complaint
- Company public response: The response by the Company to the complaint
- Company State: State where the Company is
- ZIP code: Zip Code of the Company
- Tags: Tags
- Consumer consent provided?: If the Consumer consent his Issue to be shared
- Submitted via: How the Complaint was sent
- Date sent to company: The date The company recieved the complaint
- Company response to consumer: Company's response to the complaint

*I finished this project indivisually

Author's address: Jaskirat Singh Bhatia, j6bhatia@uwaterloo.ca, University Of Waterloo, University Ave West, Waterloo, Ontario, N2L 6G9.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Association for Computing Machinery.
Manuscript submitted to ACM

- Timely response?: If the Company provided an on-time response to the consumer
- Consumer disputed?: Was the Consumer Disputed
- Complaint ID: The ID of the Complaint

2.2 Submitting Categories

There were six ways how the data complaints were sent.

- Email
- Fax
- Phone
- Postal mail
- Referral
- Web

2.3 Cleaning The Data

Cleaning The data was a hectic job as It was just dumped into the file. There was a lot of missing data and also a lot of random data which had to be removed(It gave me a lot of errors). The missing data was not removed, but just replaced by NA or left blank so that we did not lose data(as only a few columns were missing and we could use the rest). Approximately 1/6 of the data was missing. After cleaning the data, I reformatted it so that it could be read by spark accurately. Dealing with real data is very hard.

3 METHODOLOGY

I used Scala / Spark to do all the heavy lifting of the project then used R to perform visualisations.

All the Cleaning was performed on spark To get the data for visualisation, I did RDD transformations to get the required data and then stored it in different text files. The size of each text file was approximately 2kb.

For Ex: If i had to get the total number of complaints under each product, I first selected only the products and mapped each of them to 1. Then a simple map-reduce job with product as the key gave me the total products(under each category). Then to visualize it, I drew a bar graph where each bar was a product category and the length of the bar was directly proportional to its total.

4 RESULTS

4.1 Data Groups

The data was divided into four sub-groups:

- Group 1: For total(total values or count) less than 100
- Group 2: For total less than 1000
- Group 3: For total less than 10000
- Group 4: For total less than 100000

These sub-groups were made so that the data could be visualised easily.

Inside the sub-groups, if the width of the graph is thicker, that means that the total of that graph is higher.

4.2 Products

There are a total of 18 Products:

- Credit reporting, credit repair services, or other personal consumer reports
- Money transfer, virtual currency, or money service
- Payday loan, title loan, or personal loan
- Bank account or service
- Checking or savings account
- Consumer Loan
- Credit card
- Credit card or prepaid card
- Credit reporting
- Debt collection
- Money transfers
- Mortgage
- Other financial service
- Payday loan
- Prepaid card
- Student loan
- Vehicle loan or lease
- Virtual currency

4.3 States

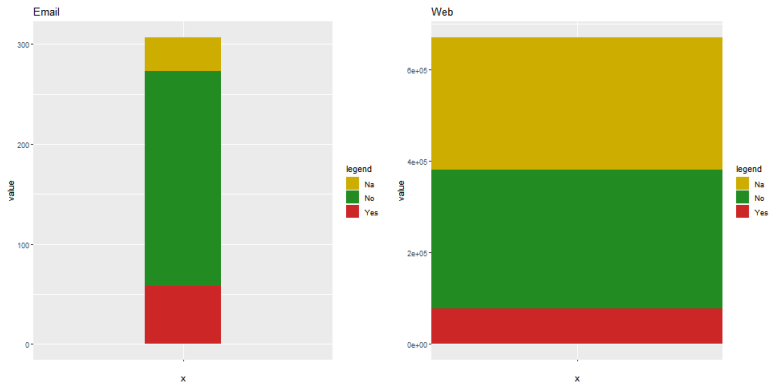
There were a total of 64 states plus there was also missing data for the State name(No name for the state provided).

The state name was a 2 lettered code. Ex: "AA" except for one where the entire name was mentioned : "UNITED STATES MINOR OUTLYING ISLANDS"

4.4 Disputed Customers

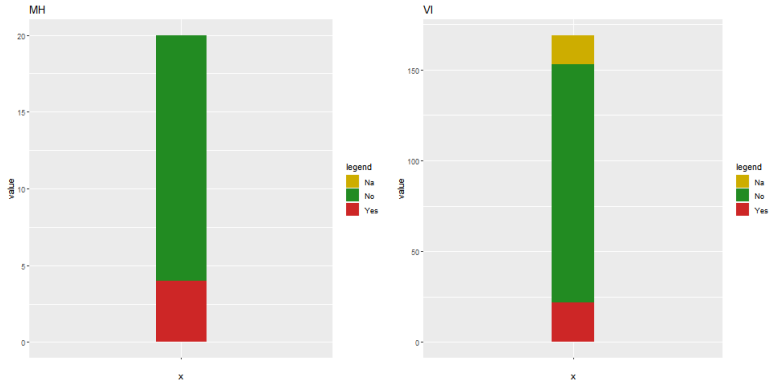
4.4.1 Overall.

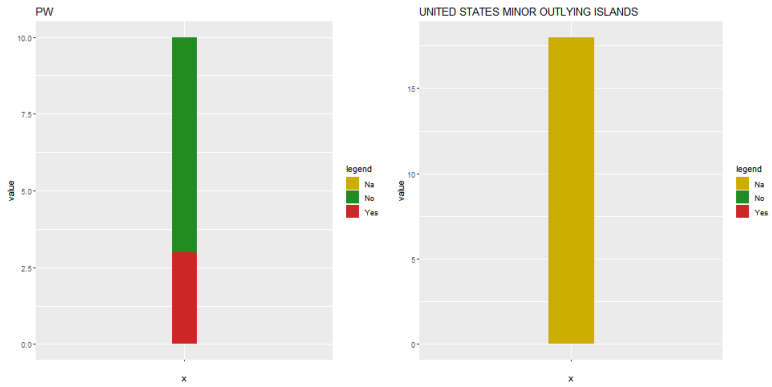
- Web was the most used of them all
- Mail was the least used of them all
- Mail had the least ratio for number of Disputed Customers
- Mail surprisingly also had the most ratio of disputed Customers
- This can be because Mail was in Group 2 and had the least number of missing data.(The total in group 2 is significantly less than that of group 4) and rest all the categories belonged to group 4)



4.4.2 *State wise.*

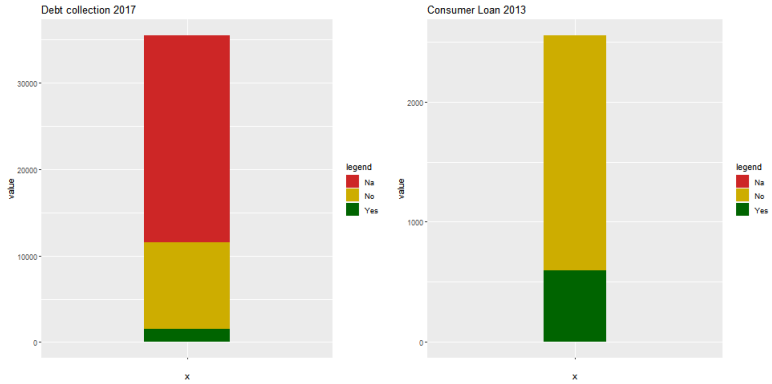
- They all followed a similar pattern - Roughly 15% were Disputed, 55% were not disputed and the rest was missing data
- "PW" and "MH" had no missing data but it also belonged to group 1. ie, very less oppertunity to miss data
- "VI" and "MH" had the least ratio for number of disputed customers. They belonged to group 2 ad 1 respectively, which are small groups.
- All the data in "UNITED STATES MINOR OUTLYING ISLANDS" was missing.
- The data where the name of the state itself was missing contained the maximum amount of missing data
- Web





4.5 Checking Customer Dispute Under Each Product Yearly

- Missing Data for: Checking or savings account in all years
- Missing Data for: Credit card or prepaid card in all years
- Missing Data for: Money transfer, virtual currency, or money service in all years
- If we look into each Product, the ratio of disputed customers decrease each year(Which is very good)
- The ratio of Disputed customers in The Product "Consumer Loan" is very high in the years 2013 and 2014, then slowly reduces after 2014.
- Missing Data for: Debt collection since mid 2017 to 2019. It makes no sense why they stopped collecting data for this Product or, maybe a law was passed mid 2017 where posting Debt data was no longer legal.



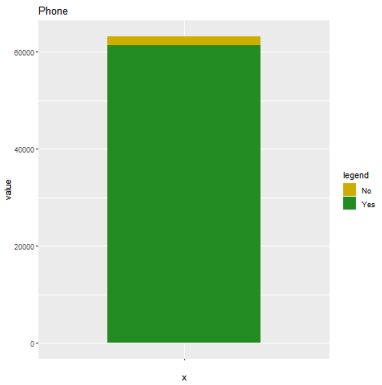
4.6 Did the customer get a timely response

Surprisingly, There was no missing data in this column.

4.6.1 By How the complaint was submitted.

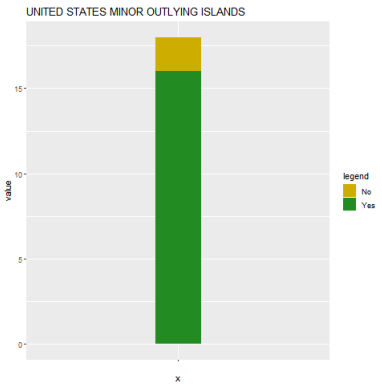
- All the categories had similar response rate
- The overall response rate was fast 96% of the time fast and the remaining 4% of the times it was slow
- Phone had the maximum ratio of slow responses

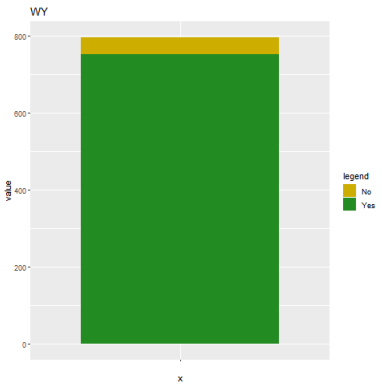
- It can be because, when we call a company, we have to hold on for a long time(They just keep asking questions press 1 for this press 2 for that) before we actually talk to a customer service agent.

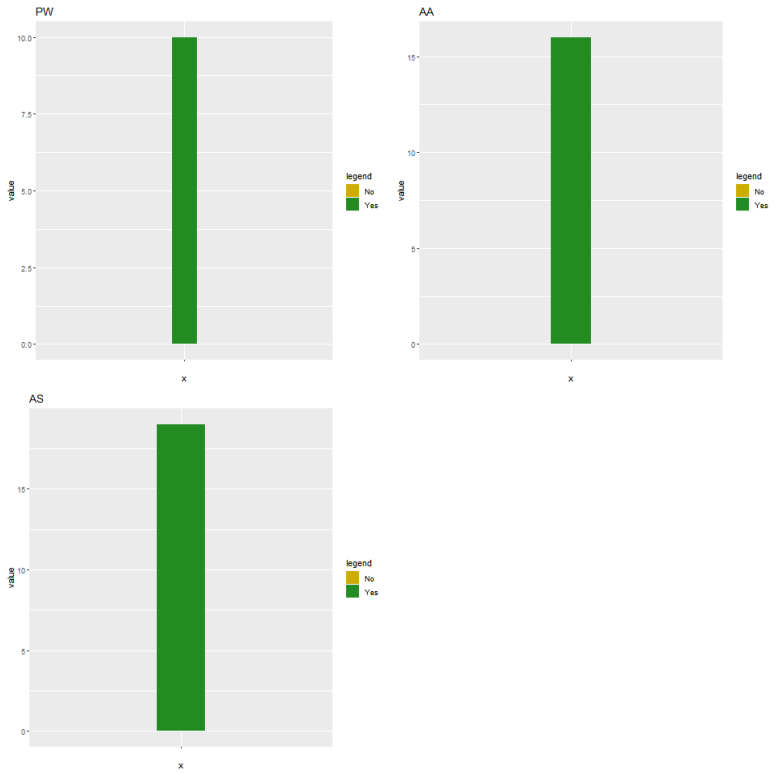


4.6.2 *State wise.*

- All the States had similar response rate
- The overall response rate was fast 96% of the time fast and the remaining 4% of the times it was slow
- "UNITED STATES MINOR OUTLYING ISLANDS" had the maximum ratio of slow responses
- "WY" was second in terms of ratio of slow responses
- "AA", "AS" and "PW" had no slow response. All belonged to group 1, so nothing major.
- Moving away from group 1, "NY" had the least ratio of slow responses
- For the data which had no State name, It followed a similar trend as the overall

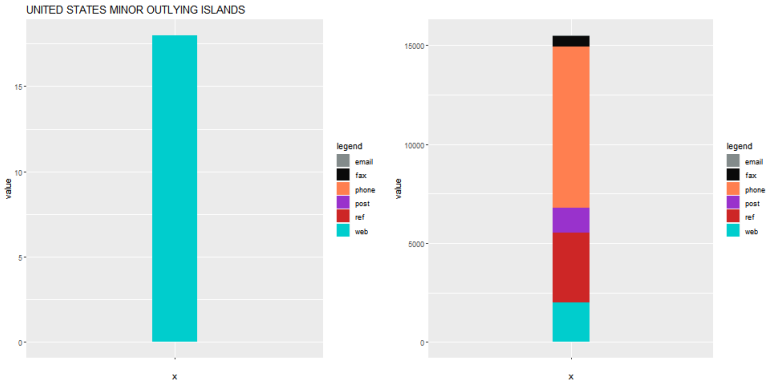






4.7 How Complaints were submitted in each State

- No missing data for this, there was a category provided all the times.
- It followed a general trend where Web was the most used, then came Referral, Phone, Fax, and then Email.
- People in "UNITED STATES MINOR OUTLYING ISLANDS" used only Web
- The data where the state was missing, there were abnormally many cases where the category was Phone(The missing state data belonged to group 4 which means it is very significant).
- This can mean that the customer service agents maybe use a landline and can not see the first three digits to determine the area code of the caller. Or that the caller is generally Disputed(which is not the case). The third possibility being, that the agent does not ask where the customer is from.


UNITED STATES MINOR OUTLYING ISLANDS
value
15
10
5
0
x
legend
email
fax
phone
post
ref
web
value
15000
10000
5000
0
x
legend
email
fax
phone
post
ref
web