

PROCESS BOOK

Visualizing US Consumer Complaints Data in Finance Sector

GitHub Repo : <https://github.com/madhur12/Visualizing-Consumer-Complaints>

Members :

- | | | |
|------------------|---------------|---|
| 1. Shlok Patel | Uid: u1083432 | Email: shlokipatel@gmail.com |
| 2. Madhur Pandey | Uid: u1065393 | Email: madhur@cs.utah.edu |

Background and Motivation:

In today's era, everyone uses one or the other kind of financial products/services offered by various financial institutions. Almost everyone has a bank account or a credit card, while some others might also have loan related products. Therefore, it becomes necessary to analyze how satisfied people are with such services and which institutions are doing their best in providing these services to their customers. It would also help if we could analyze how the complaints statistics have changed over the years. So, we decided to choose this project so that it can help us take better decisions in the future, when dealing with finance sector.

Project Objectives:

The main objective of the project is to gain meaningful insights from raw data collected by CFPB, so it can benefit both the consumers and providers of financial products/services.

The primary questions the visualization is trying to answer includes:

- Major issues faced by the consumers when using financial products/services.
- The number of complaints responded/resolved.
- Best/Worst performing institutions based on various criterias
- Regional complaints pattern in the USA.

For the consumers, the visualization can help in determining which institutions are best suited for the service they need.

For the institutions, checking these trends can help to improve their products/services and check how they are performing compared to their toughest competitors.

Data:

The Consumer Financial Protection Bureau (CFPB), a Federal agency responsible from protecting the rights of the consumers, is collecting the consumer complaints data we are using since 2012. The data is available in .csv, .json and .xml formats. All the data is available on the data.gov website at the below link:

https://catalog.data.gov/dataset/consumer-complaint-database#topic=consumer_navigation

INITIAL PROPOSAL

Data Processing:

We might need minimal data processing. As of now, we only need to derive new date columns from the existing ones.

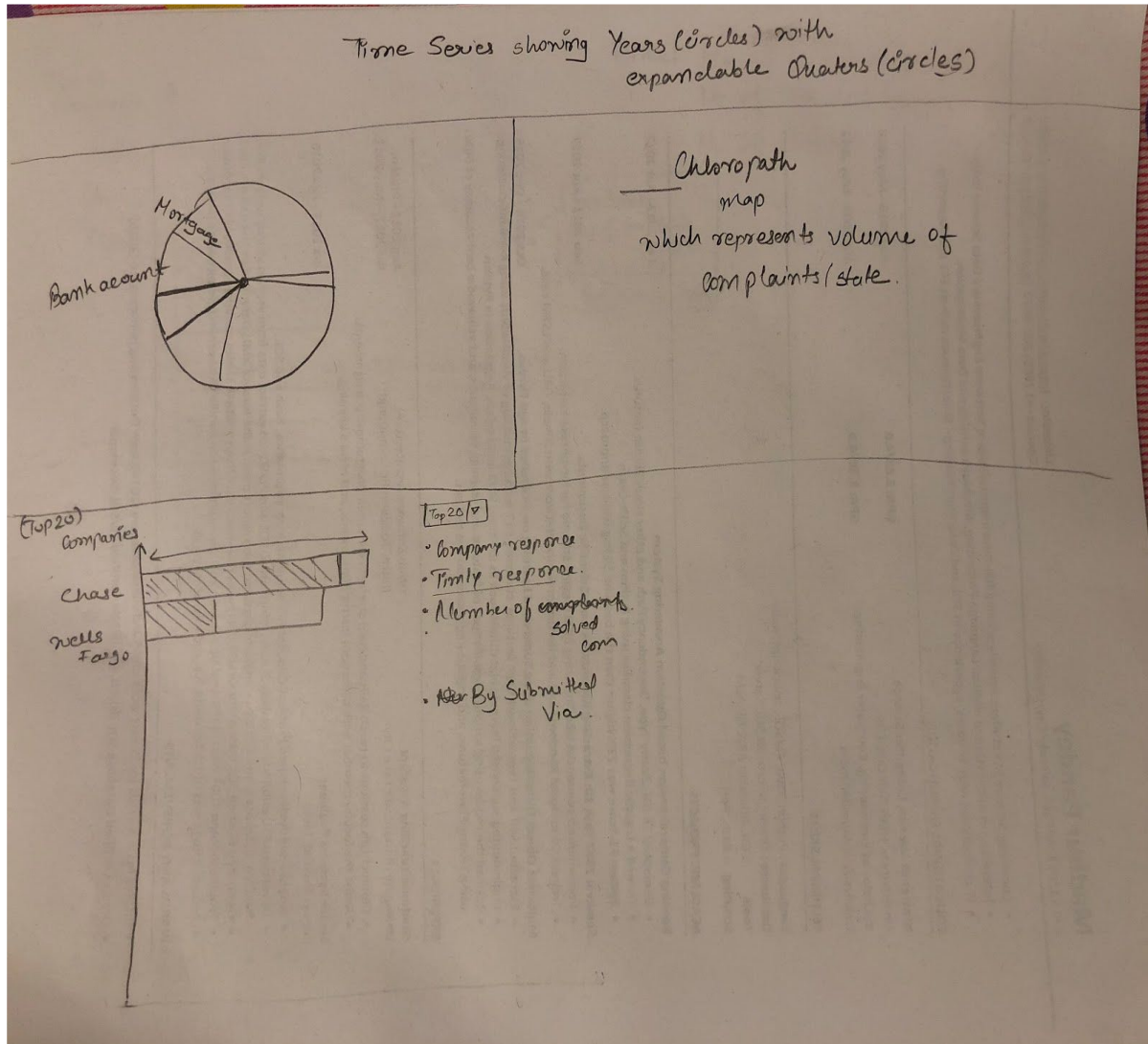
The dataset is very huge and we might have to select a random subset of the data so that it is faster in processing and represents the entire data. However, this is tentative.

Visualization Design:

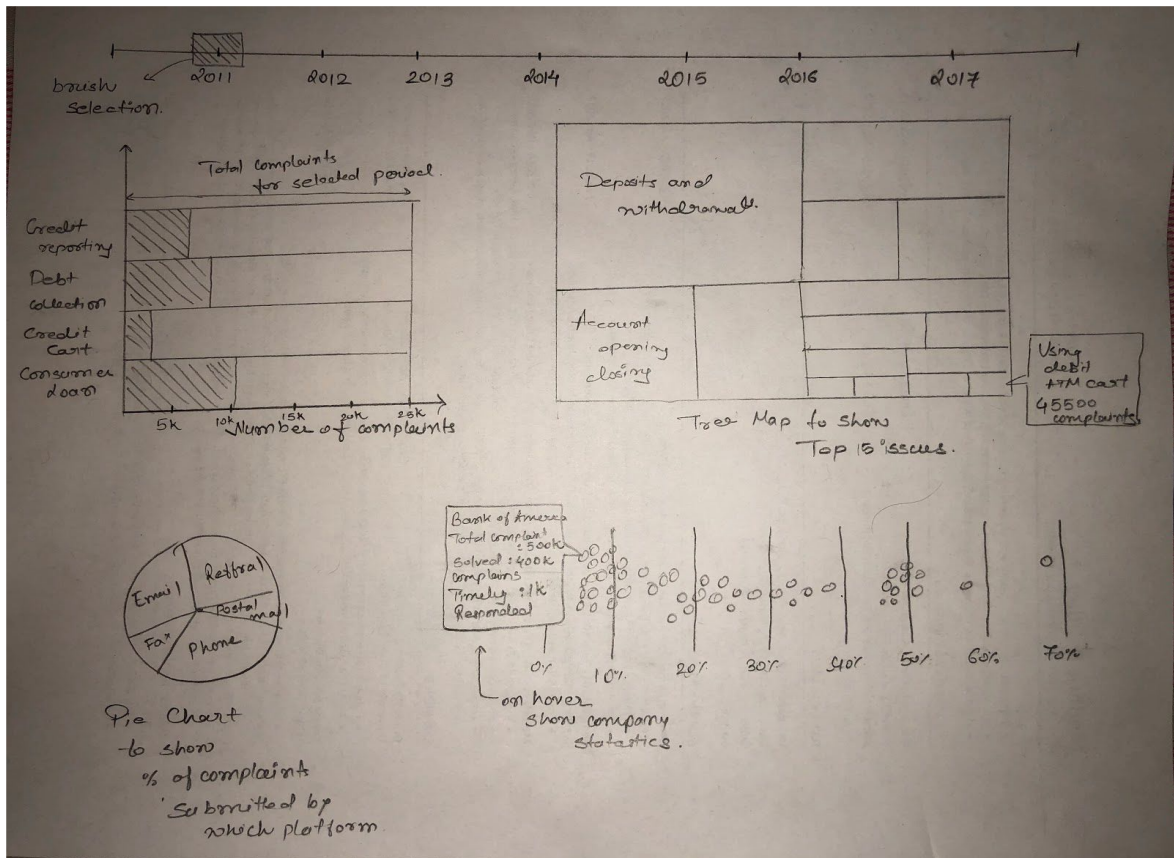
Our visualization design started with brainstorming the various questions we want our visualization to answer. After finalizing the questions, we came up with various initial designs and finally agreed upon a final design. Below are the snapshots of our initial designs and final design.

Initial Designs:

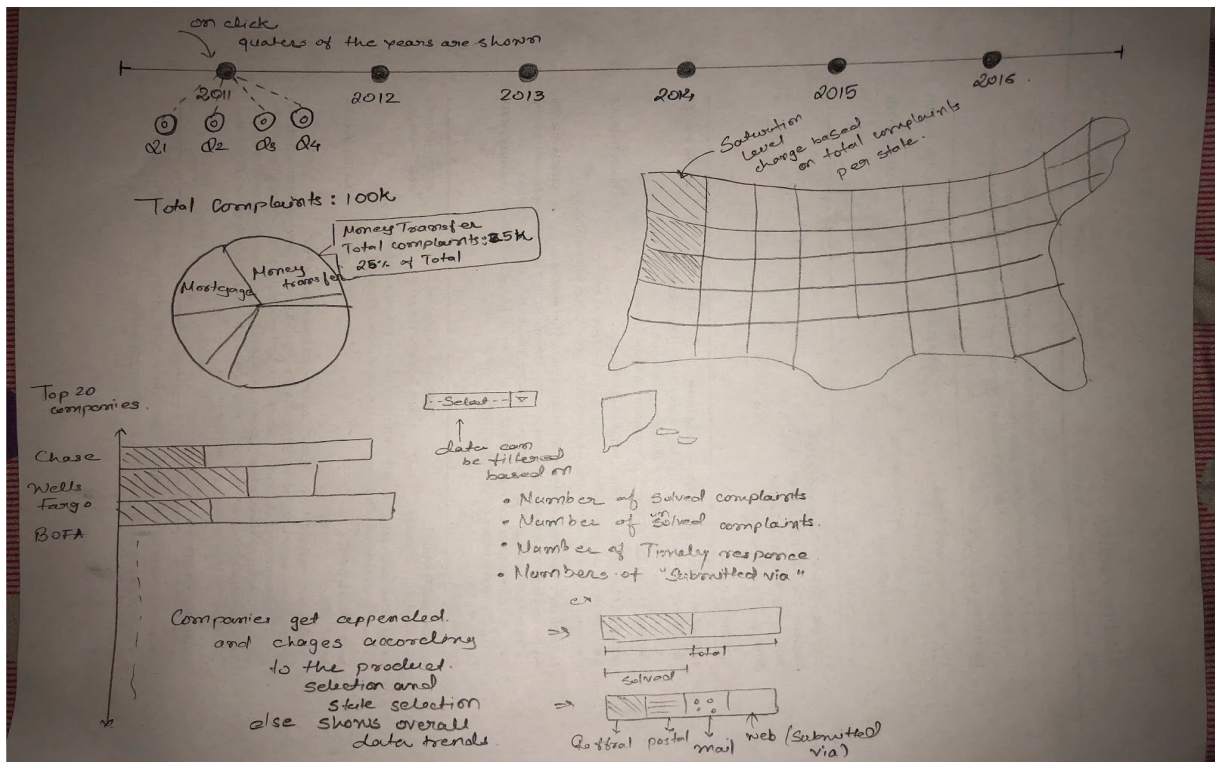
1.



2.



3:



Final Design:

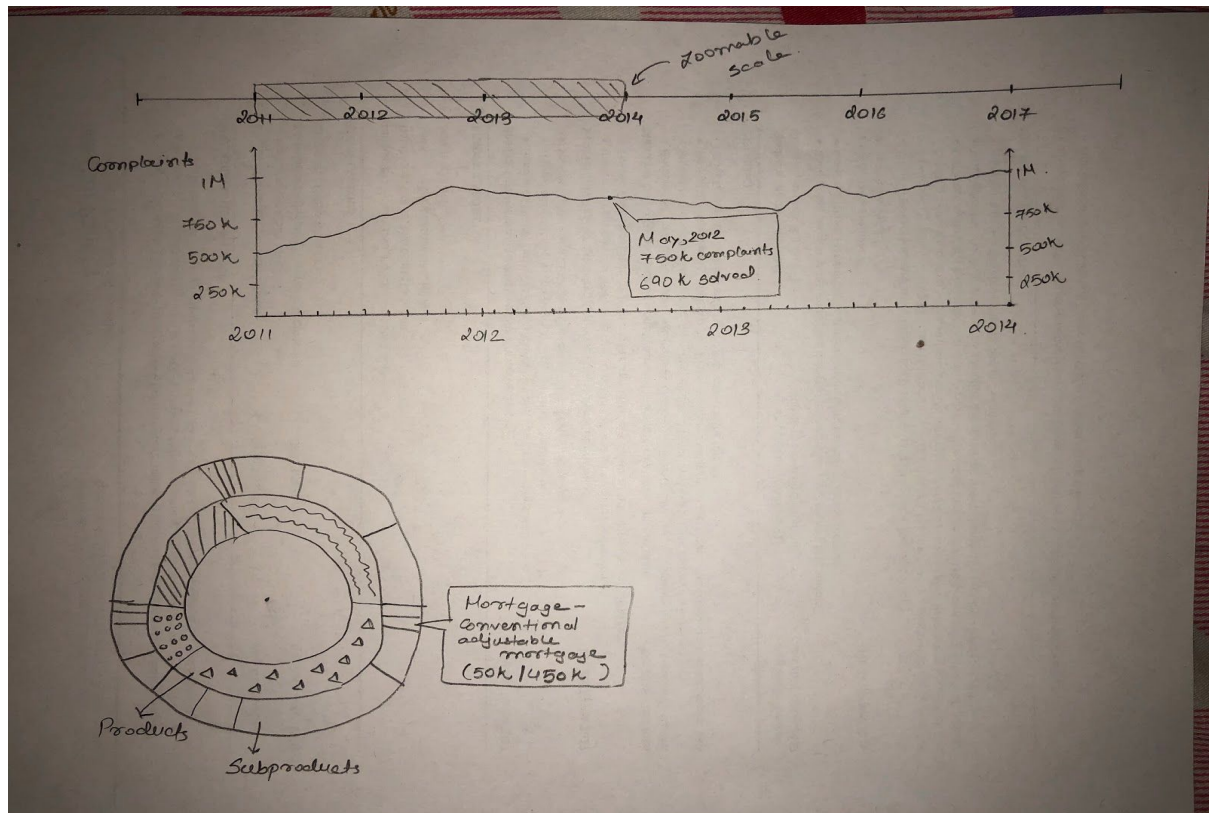


Fig 1: Complaints Trend over year and Product/Subproduct Sunburst Chart

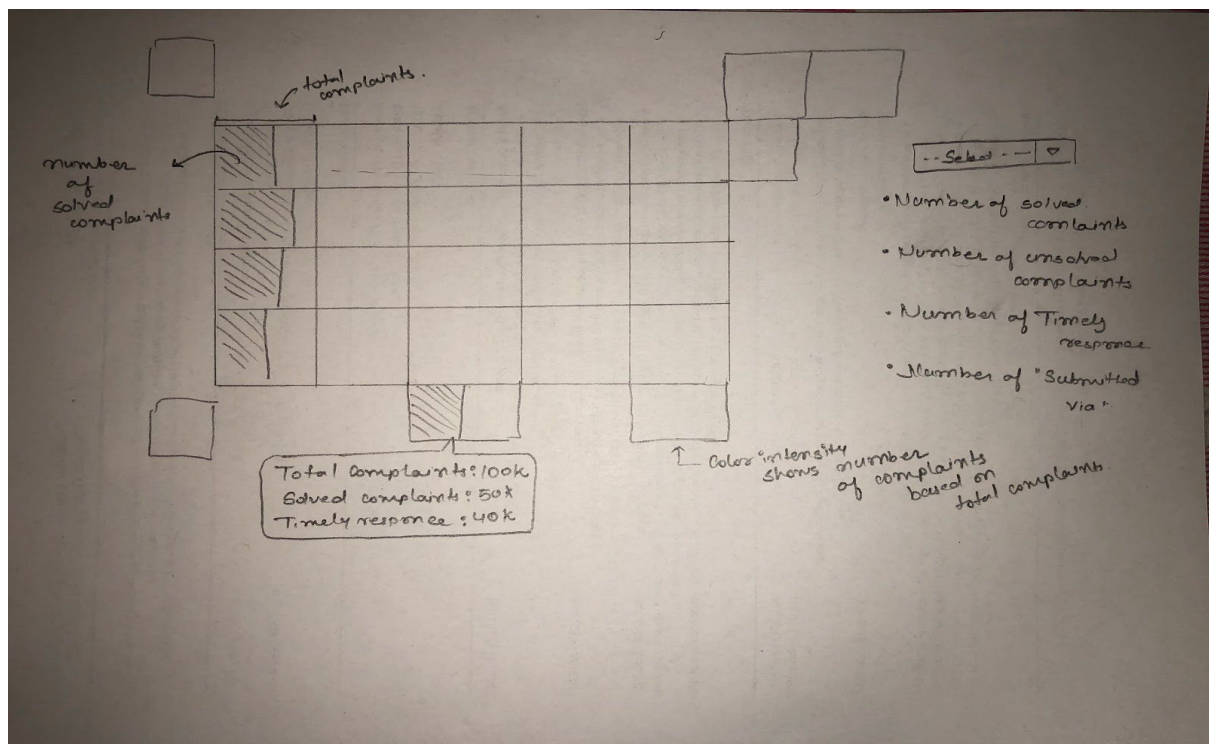


Fig 2: Map with area charts representing the overall state trend for the dropdown selected.

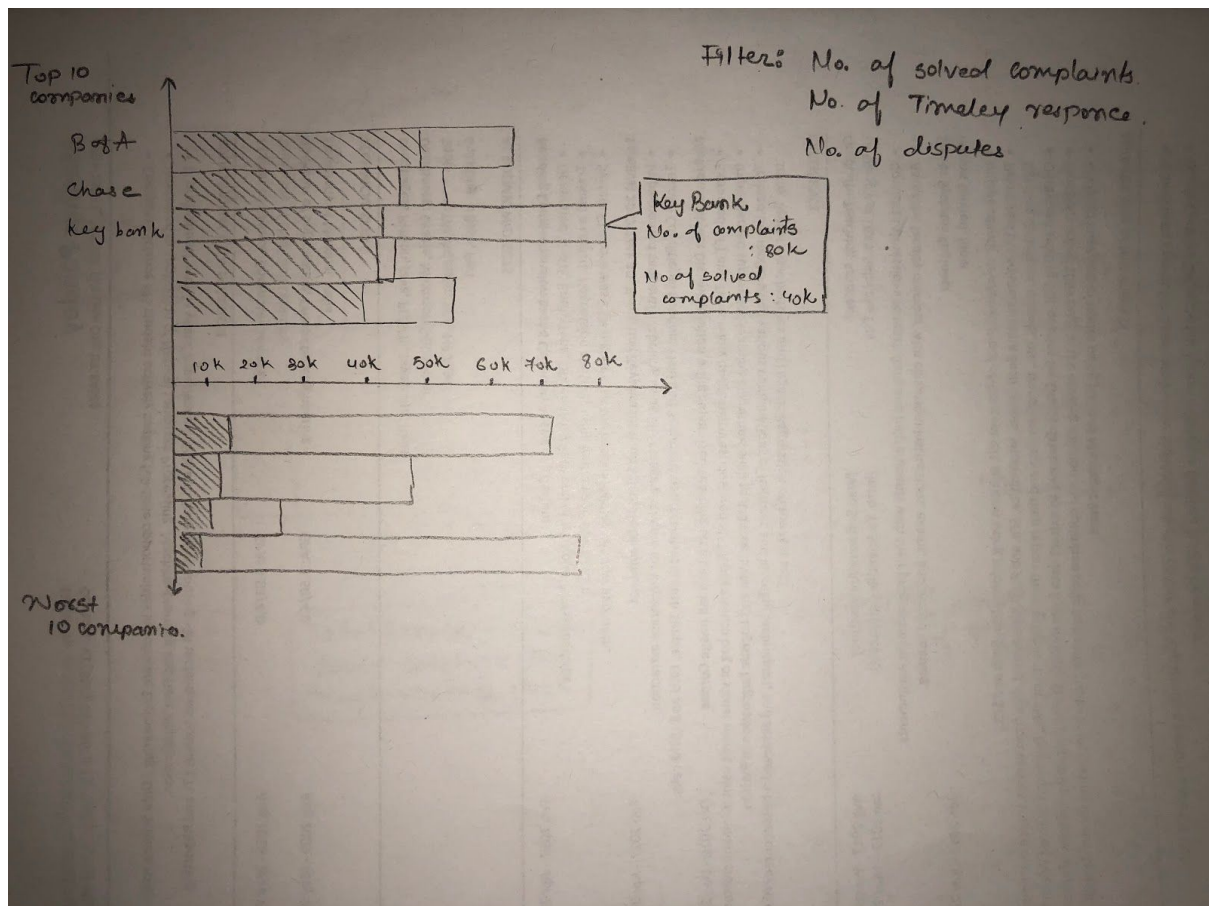


Fig 3: Stacked bar chart showing best/worst companies based on the dropdown value selected.

We plan to implement the visualization using time series, zoomable sunburst chart, customized choropleth and stacked bar chart. We will have brush component over the time series which will be used to select a subset of time and change all the views accordingly. The subset can also be selected by zooming over the timeline.

All the svg components will have tooltips which will display additional information about the data it visualizes. We will also have drop-down for the stacked bar charts to compare the best/worst institutions based on different criteria.

All our visualizations will be interactive in the way that selecting something in one view will update all the other views accordingly. All our views will be interactive in the way that selecting something in one view will update all the corresponding views.

The various channels used are:

Position:

The position on a common scale is used as the channel to encode the year time line. The position channel is the best visual encoding for Ordinal type.

Position/Length:

The complaints trend and the best/worst company charts use position/length as the visual encoding as both the fields are quantitative in nature and quantitative data is best distinguished using the position/length channel.

Hue:

The Sunburst chart uses hue as the channel to distinguish between the various products and subproducts. This is because the products and subproducts are categorical data and hue is a good separator for categorical data.

Arc Length:

The Sunburst chart also uses the Arc length as the channel to encode the number of complaints for each product category, which is quantitative data.

Area:

For each state on the map, we are using Area as the channel to show the proportion of total complaints received vs resolved.

The various marks that our visualization will be using include:

Line:

Lines are used to show connectedness in complaints trend chart. Along with the position channel, it helps in better visualizing the pattern in the chart.

Saturation:

We might (tentative) use saturation to encode the volume of complaints received in each state. This is the saturation of the area chart for map described above.

Must-Have features:

The interactive complaints trend chart and the best/worst companies stacked bar chart, which change based on brush selection are must have feature.

Optional Features:

The Sunburst chart is an optional feature. Also, if time permits, we will implement all companies response time plot as drawn in the bottom right of design-2.

TIMELINE

Oct 16,2017: Project Team Meeting 1

Today we had a review session with each other. We had divided the work and had decided to check with each other on weekly basis. We decided to split the work and first make each view completely and later on in the project we will meet and make these views interactive with each other. This would probably make the process faster as we would not depend on each other's work. For the first week we decided to split the Timeline chart and the stacked bar chart amongst us as they were the mandatory components of the project without which the project is a failure. Also some time was given to clean the data and merge new columns which would be used for plotting the coordinates of the map.

Oct 19, 2017: Repository Setup

Madhur created a Github repository and added Shlok Patel as a collaborator for the project. Initial Project Proposal submitted via the same.

Nov 3, 2017: Project Review

The major critical feedbacks that we received from the individual team members of the reviewing team are as below:

1. The overview visualization doesn't seem to be specific to the target audience and improving the same might help in getting better insights from the visualization.
2. Having a color scale in map (heat map) instead of bar chart, as bar chart might make the map look messy.
3. Customer might want to check the performance of an individual company. So, adding a filter on companies to highlight states and see how the company is performing can be very helpful.
4. Customers would like to compare the performance of 2 companies.
5. Moving between the views instead of having everything on a single page.

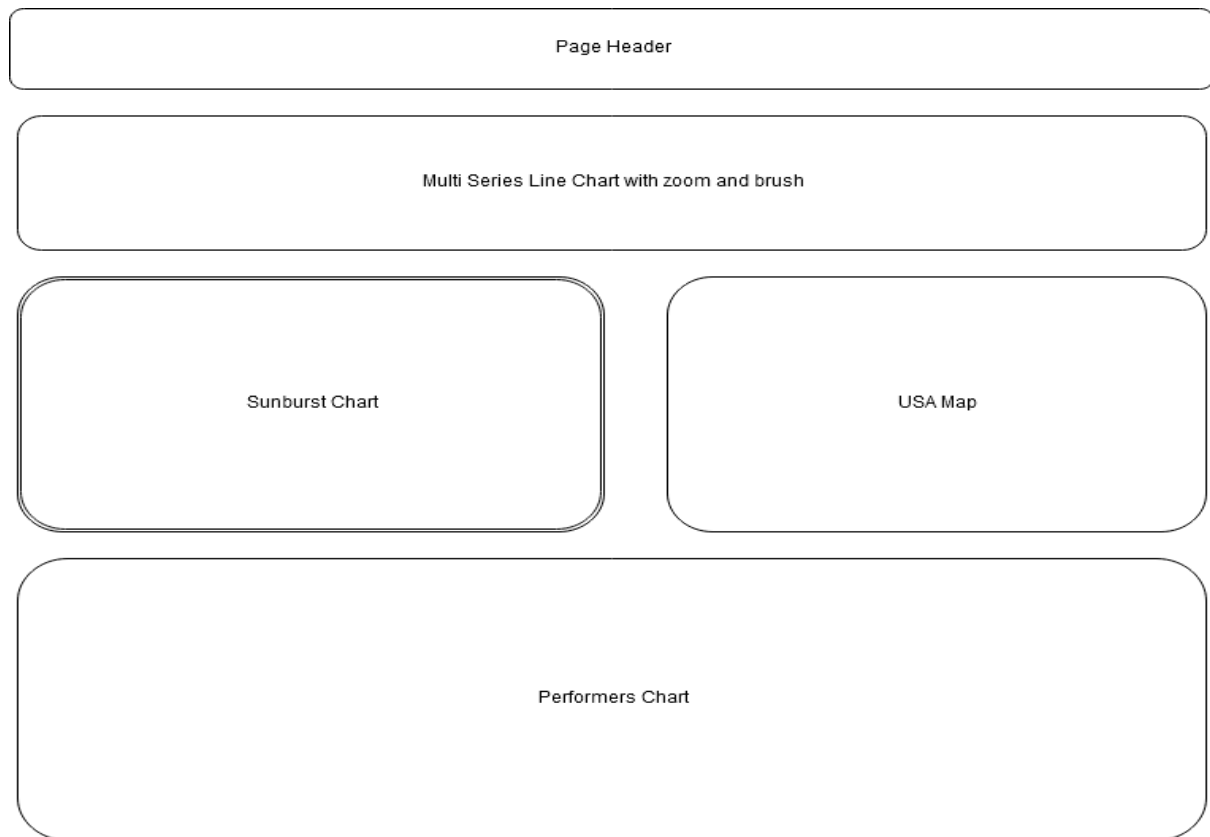
Nov 5,2017 : Project Team Meeting 2

This was the second project meeting and we had almost finished cleaning the data and also make the basic structure for the Timeline chart and the stacked bar chart. Still much improvement is needed in the stacked bar chart as it is still in the initial phase of development. We discussed some problems which we could face in the future related to the design which was basically related to the interaction between different views. As we were using the zoomable brush we were not sure on how to implement those things. We are still not sure of that and we are planning to change the approach a bit.

Nov 8, 2017: Data Analysis and Work Distribution

Both the team members meet up to further decide the approach to be used towards the implementation and check for any progress from any team members. Until now none of the member was able to put in dedicated hours towards the project.

It was decided in this meeting how the structure of the html file would be. We came up with the below structure for our webpage.



We then tried analyzing the data using excel. However, due to the large size (~400MB) of the csv file, we were unable to make much progress. We then turned towards Linux to truncate the file before using, however, due to newlines in the file, the truncation resulted in a file which was no longer in a valid csv format.

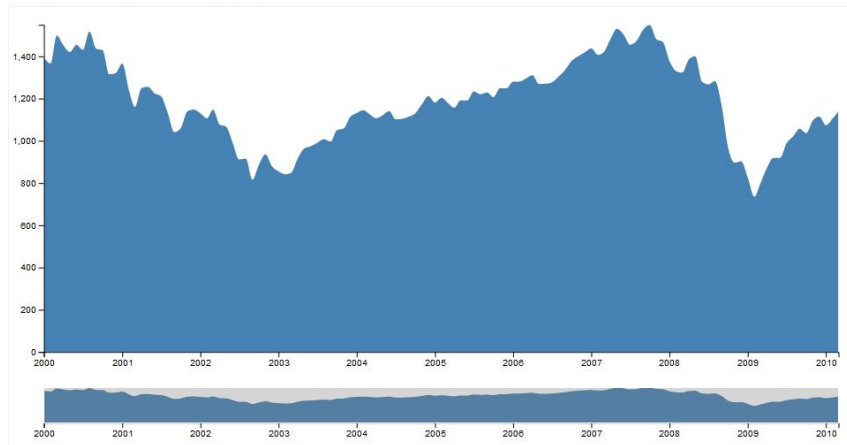
Finally, we decided to remove fields like "Complaint Note" and "Company Response Note" (the ones containing newlines) from the data as they were never used in the visualization and did not have any significant impact even if included. This reduced the file size significantly to 150 MB. Finally, the file was still about 150 MB which took a while to load in D3.

It was also decided in this meeting that Madhur would work on the Timeline chart and Shlok would take care of deriving additional fields in the file for Maps, as well as work towards the performers chart.

Nov 9, 2017: Data Size-Still an issue

Madhur worked on the initial version of the Line Chart as decided. The below example from Mike Bostock's blocks were taken as reference while working on the zoomable and brushable line chart.

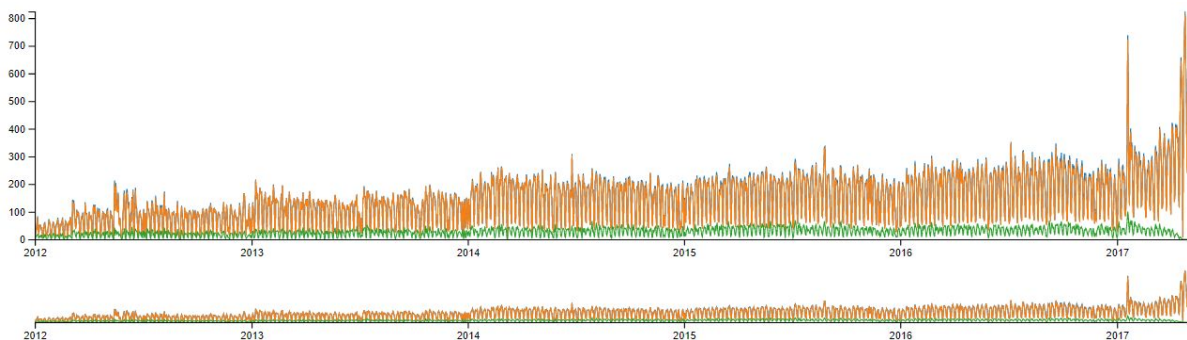
Brush & Zoom



Combining [d3-brush](#) and [d3-zoom](#) to implement Focus + Context. Another approach is to [zoom to the brushed region](#).

[Open](#)

Subsequently, the initial version of the Line chart looked as below:



Today, while working on the Line Chart, Madhur realized that loading the records was still taking a lot of time.

To overcome this problem, we decided to reduce the number of records from 800k to 300k just for the sake of faster page load.

We checked the outcome of the line chart before and after truncating the csv file. We were satisfied that the truncated data was still a good representation of the overall data.

However, the initial version of the line chart still looked a little messy. And the zoom and brush feature are still not implemented.

The format of the input file was changed for the specific implementation of the timeline line chart. The data was nested and rolled up with key being the date received and count of "Timely Response", "Disputed Field" and were also calculating.

This data is used to create a timeline where the y axis shows the total complaints, Complaints which are Timely Responded and complaints which are disputed and the X axis is the Date in the Form of months or year based on the selection of the brush and the zoom.

At the same time, for creating the map Shlok included the fields "row" and "column" of the particular state which would be helpful to create the Map.

As of now, for different charts we are creating a custom data structure (JSON) which includes all the fields necessary for the making the visualization. Later on we will pass the corresponding data to other charts to make it interactive.

Nov 10, 2017:

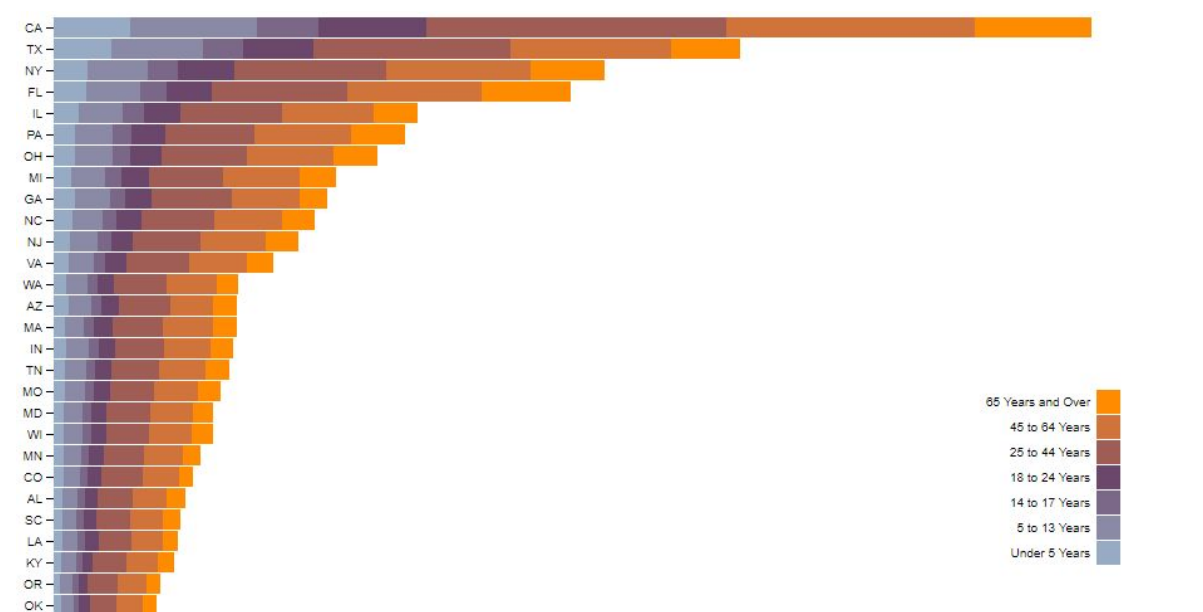
For the PerformanceChart.js, Shlok created the JSON object using nest() which includes "Company" field as key and "Timely Response", "Disputed Field", "Submitted Via" and we are also calculating the total number of complaints for particular company and including it in the data structure as a "Total" field

This data is used to create a stacked bar chart where the y axis shows the company names and the X axis is the shows the total complaints against Complaints which are Timely Responded and complaints which are disputed.

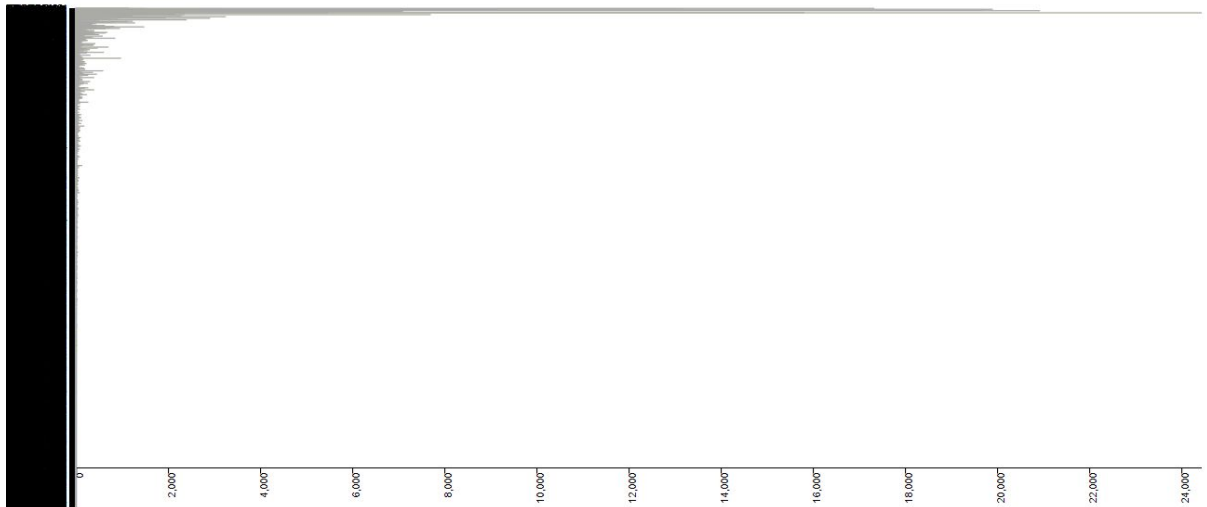
As of now we will display all the companies over the Y axis. Later on in the project we will filter the companies based in the selection form the other charts and just display the top 10 performers and the worst 10 companies and their data.

Initial Implementation

Shlok used the below stacked bar chart is the reference. The stacked bar chart design would look as shown below.

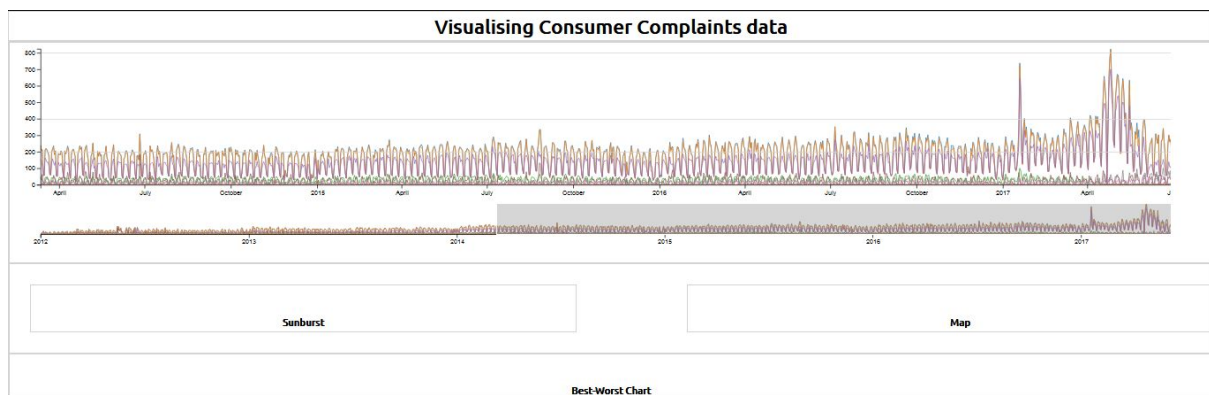


We are still working on it and its is near completion. We are currently getting 300k data over the y axis which should be sorted and we should filter only the top 10 and the worst 10 companies instead of all the companies. The initial implementation is shown below.



The next step is to make a filter for the stacked bar chart as per the final design. In the stacked bar chart the total complaints will remain as it is where the inner rectangle will be made based on the selection of the filter. Like disputed complaints, submitted via of the complaints and the resolved complaints.

Madhur continued working on his timeline line chart to add the zoom and brush effect. We were able to achieve the desired result as per the initial proposal. We also added gridlines and a few other information to improve the output. The webpage now look as below:



Today, we both also discussed some improvement to the above timeline chart. This includes aggregating the data as per the different time intervals based the zoom level and brush selection. We tried some approaches together, however, we were unsuccessful.

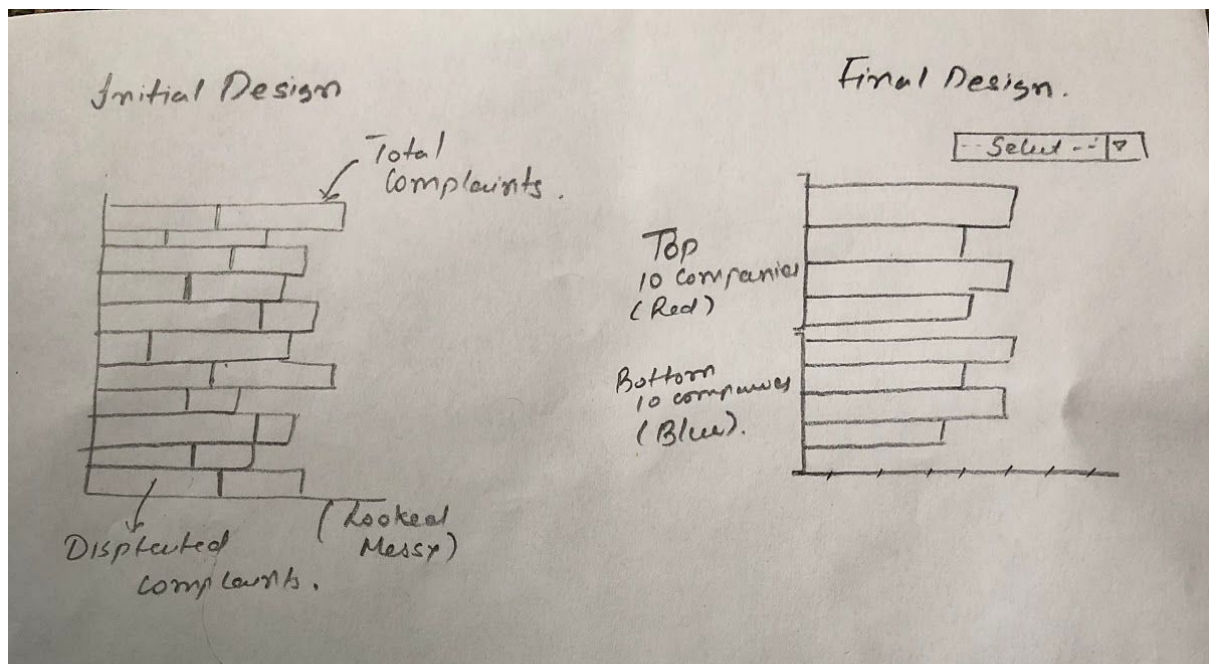
Nov 11, 2017:

Today Shlok continued to work on the stacked bar chart. And implemented the stacked bar chart based on the total Complaints as the major bar and the disputed complaints as the minor bar. The reason behind selecting this approach was to get the estimate of the difference between the total Complaints and the disputed complaints and the timely solved complaints.

But after doing the stacked bar chart for the total Complaints and Disputed complaints, we discussed about it and came to a conclusion that the stacked bar chart just gives an estimate of the total Complaints and getting the top and bottom 10 companies does not give the information that the consumers are looking for.

Total Complaints are not the correct criteria to judge a financial sector company. Because the company may be good and can be so huge that it has more number of complaints than a smaller company with less customers. The criteria which really matters is the percentage of the disputed complaints over the total complaints and the percentage of the timely solved complaints over the total complaints.

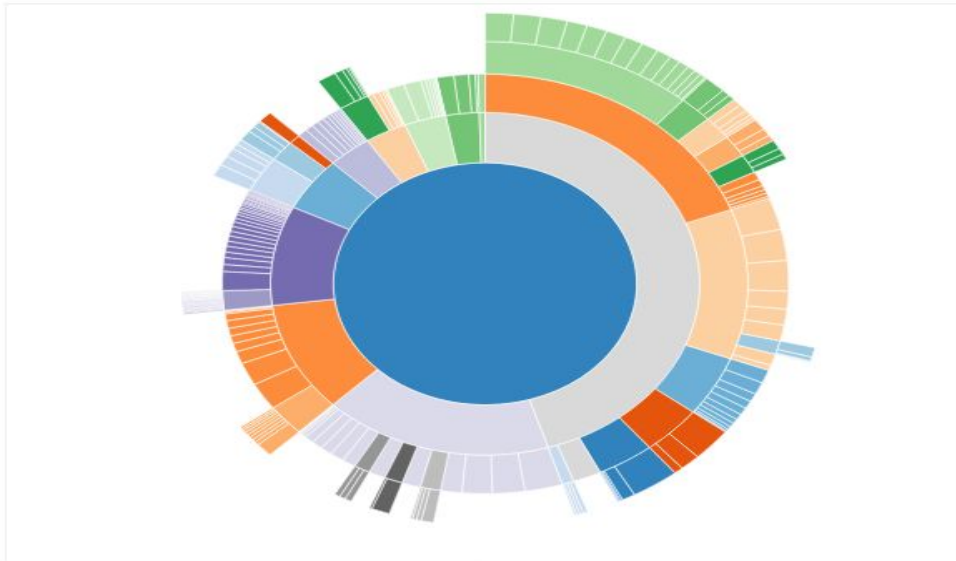
So it totally depends on the consumers which criteria they are looking for. It can be total complaints, total disputed complaints, total timely solved complaints, percentage of the disputed complaints over the total complaints and the percentage of the timely solved complaints over the total complaints. So we decided to place a filter at the top of the performance chart (bar chart) which shows all this filters as the drop down and on selection of the filter the bar chart will get populated with the data.



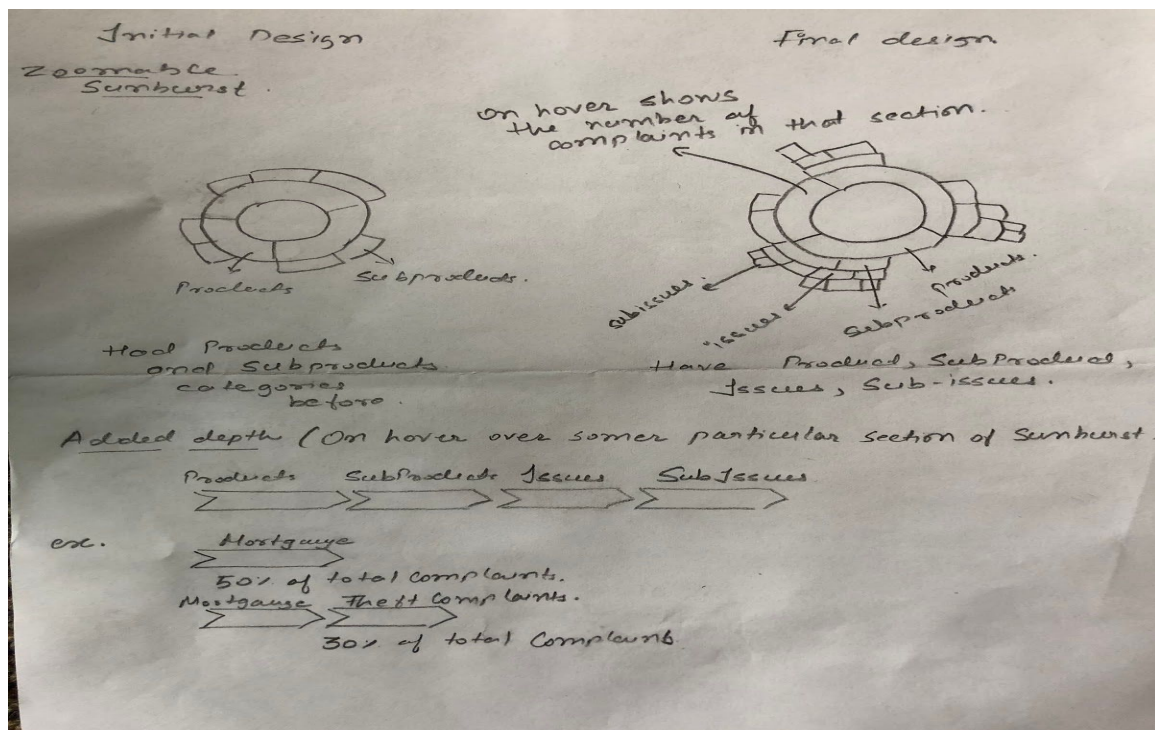
Nov 13, 2017: Implementing the Sunburst

Madhur started with the implementation of Zoomable Sunburst. The initial idea was to implement Mike Bostock's Sunburst chart as below :

Zoomable Sunburst



It took a total of 3 days to get the zoomable sunburst completely running. Once the implementation was complete, we found that the User would have no idea about his traversal path. So, we decided to add features to our Sunburst. Below is the design of the navigation trail that we decided on the meetup.



In the meantime, Shlok completed implementing the bar chart which change based on the selected Comparison Category. The best companies were colored in blue and the bottom companies were colored in red which is the general trend showing the top and bottom companies.

There might be other modifications needed in the chart based on the selected data from the timeline and the sunburst as they also apply the filter like time and different categories in the sunburst like product subproduct, issues and sub issues. Depending on all this data the bar chart will get populated.

If only 2 companies were satisfying the criterion then they were just shown on the top of the bar chart indicating that both are the best companies. If there were 7 companies then it shows top 5 as best and 2 as the bottom companies which is not what we want and so we had to change the approach to handle this kind of scenarios.

Nov 15, 2017: Project Team Meeting #3

We met for the third time and checked the project status and reviewed each others approach. A few more important changes design changes to improve user interactivity were decided.

Shlok came up with the idea of adding a Company comparison table at the bottom of the page. This would allow users to compare two or more companies. The company would be added to the comparison table by clicking on the performers chart. The next question was "What if someone have added the company by clicking on the company accidentally? Or someone don't want that company and want to remove from the compare list so that he/she is left with just few limited companies?". Most of the time its the human nature to come to a conclusion based on the elimination criteria. So we decided to also add a cancel "X" button at the end of every row so the the customer can just click on it and get rid of the unwanted companies from the list. The design would look something like below :

Added A table for company comparison.

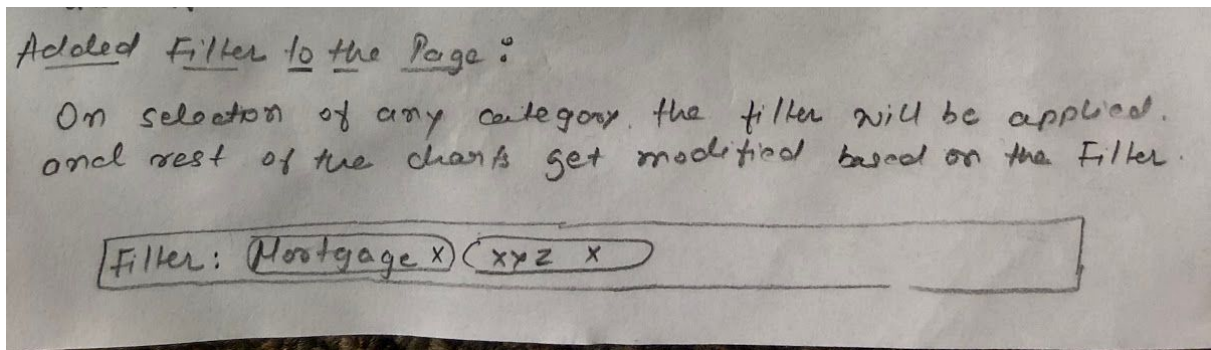
- All the selected companies from the bar chart will get populated in a table.
- Main idea. Change filters add companies to table

Company Name	Total Complaints	Disputed Complaints	Percentage Disputed	Percent Timely	
XYZ	53000	50	5%	99%	<input type="checkbox"/>
ABC	4500	100	10%	95%	<input type="checkbox"/>

Better Choice!!

We think that this helps the user to compare as many companies as they want. We can also sort the rows based on different criterias (Additional fields which were not available in performers chart).

We also thought that we need to provide the user with the idea of what the views are currently displaying. To do this, we decided to add a fixed Control bar at the bottom of the screen. The initial design of the same is as below:

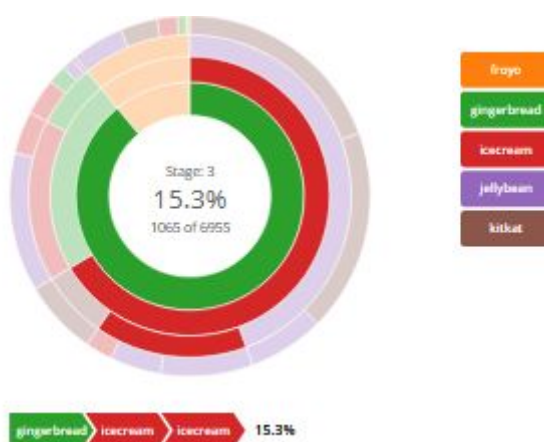


We also reviewed our codes and some coding improvements were discussed. This included how we handle the data in our individual components, what could be better transition for a given chart, and how to improve user experience using the various event handlers.

We also decided to use Bootstrap v4 for making our website more elegant.

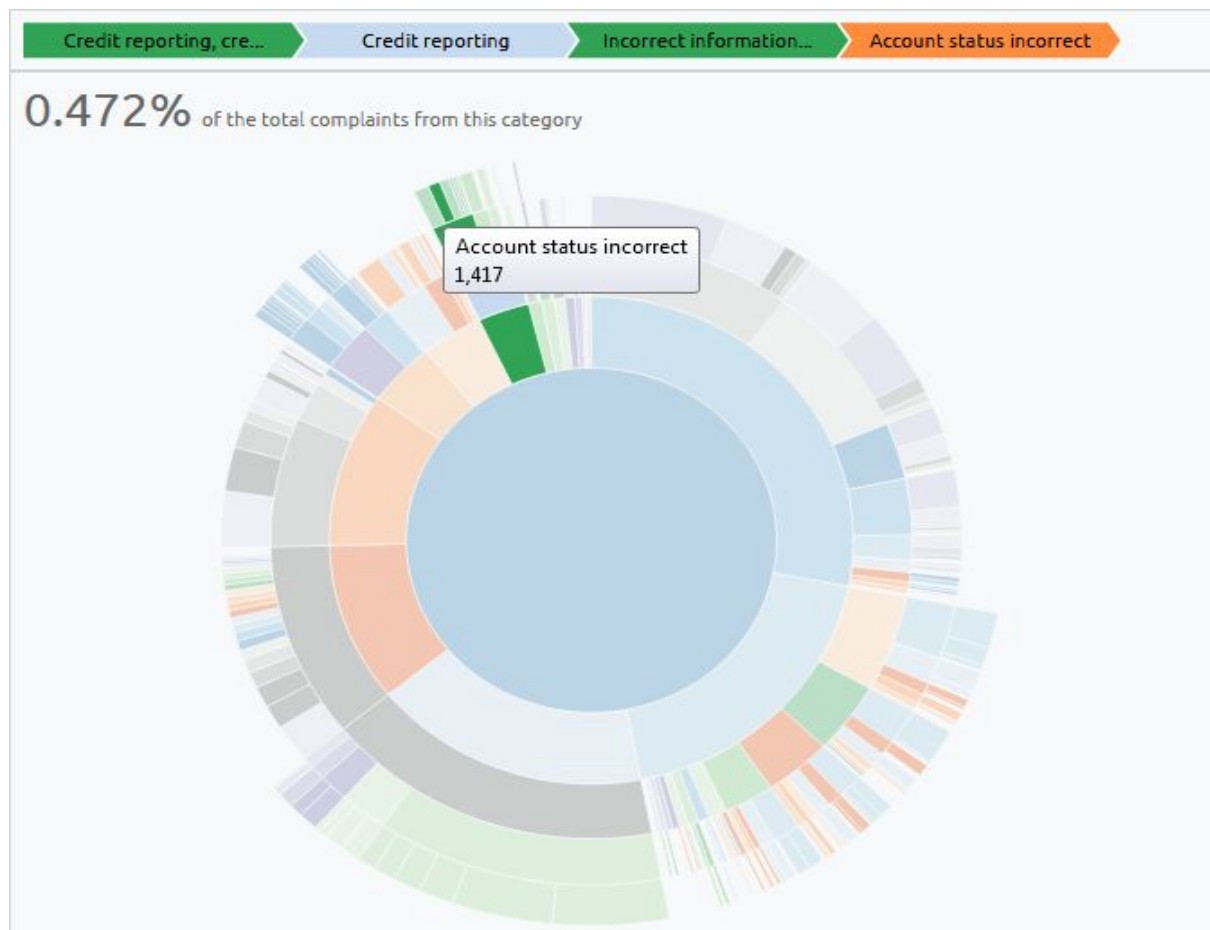
Nov 17, 2017: Final Sunburst with Trail

We referred to Chris Zou's below sequence sunburst implementation to add the trail to our Sunburst, which we decided in our earlier meeting.



The implementation shown above is not zoomable which restricts visualizing distribution in a specific subcategory. So, we planned to implement the above over our earlier Zoomable Sunburst implementation. We had to take some design decisions such as limiting the text in

the trail and displaying the percentage outside based on our view width. Our final Sunburst looks as below :



Nov 18, 2017: Generic Framework for Interactivity

It was very important to have a generic framework for interactivity. We had decided in one of our meetings that we would allow interactivity between all the views. I.e. clicking on some component in one view updates all the other views subsequently. So, having redundant and different approaches in each view would limit the interactivity between the views. Madhur came up with an approach to handle this, by having a global filter function, which updates the filters in our Control bar and also allows each view to update, depending on whether or not it needs to apply the filter.

The main idea was to have a global filters object which is updated by the views based on user interaction. We had discussions concerning the safety vs performance measure for our Visualization, due to the global scope of the filter and the object variables. However, we decided that given the size of data, redundancy could be expensive and having a global scope would be safe since all the views need to have access to all other views.

Nov 20, 2017:

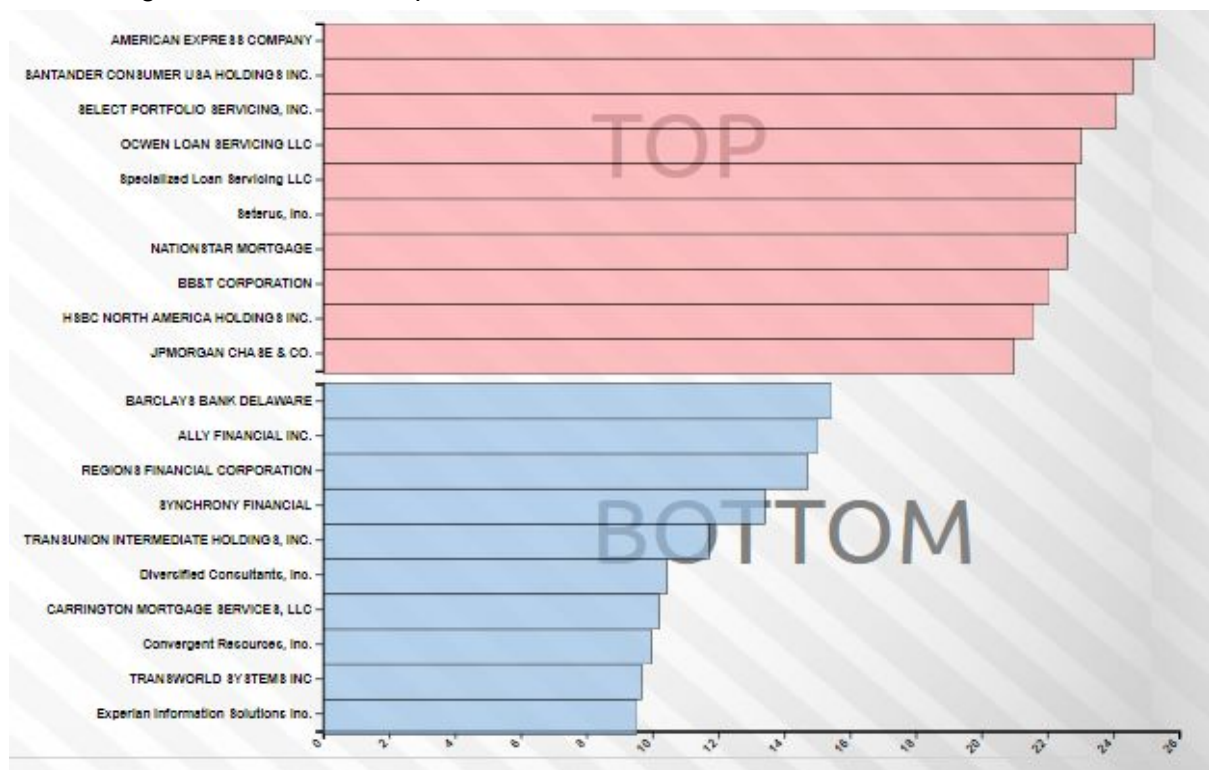
Shlok completed the review changes and interactive framework implementation for Performance bar charts. These included separating the best/worst charts and applying

transitions. The tool tip feature was also added on the bar chart which shows the exact number of the criteria which is shown by the bar chart (total complaints, total disputed complaints , total timely solved complaints .percentage of the disputed complaints over the total complaints and the percentage of the timely solved complaints over the total complaints). The tool tip helps to easily get the value and people don't have to stare at the axis for comparing the highly competitive companies with almost the same values.

The bar selection also highlights the bar to red so that user can that particular company should be added for comparison

However, it is not yet linked with the yearchart and the sunburst as Madhur was working on it and we had decided to do the first merge 2 week before the project deadline to check if the visualization as a whole.

Below image shows the final implementation of the Performance chart.



Nov 23, 2017:

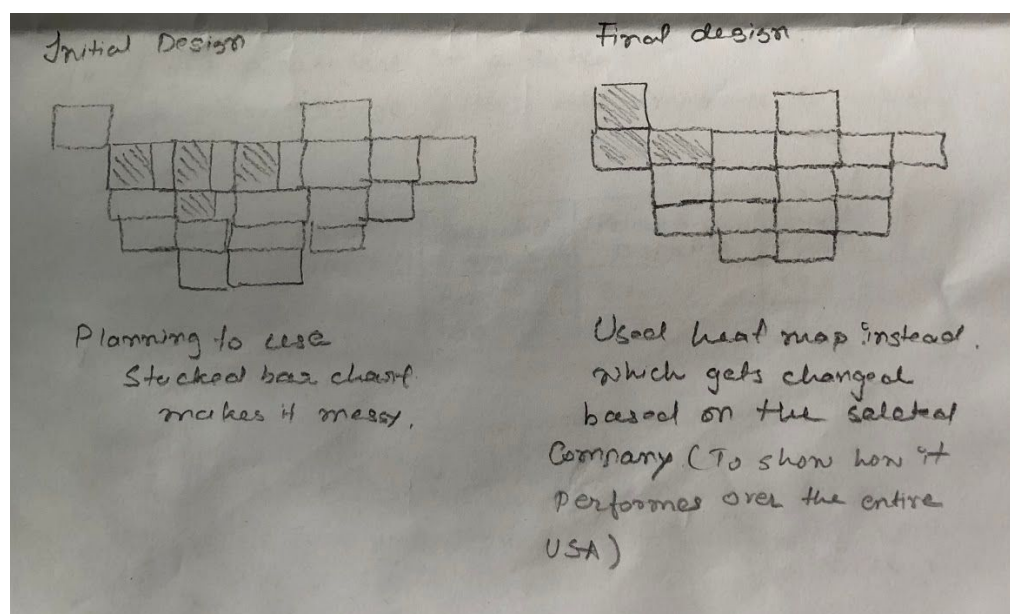
It took some time implementing the table and the cancel feature in the table. Some alternating vibrant colors were given to the table rows so that it is easy to differentiate between the two rows and it would be easy to compare the companies. Final Design is shown below

Company Name	Total Complaints	Timely Responded Complaints	Disputed Complaints	Percentage of Disputed Complaints	Percentage of Timely Responses	
Convergent Resources, Inc.	903	902	90	9.97%	99.89%	X
Diversified Consultants, Inc.	737	730	77	10.45%	99.05%	X
SYNCHRONY FINANCIAL	5464	5459	734	13.43%	99.91%	X
BARCLAYS BANK DELAWARE	1631	1629	252	15.45%	99.88%	X
TRANSWORLD SYSTEMS INC	972	972	94	9.67%	100.00%	X
REGIONS FINANCIAL CORPORATION	1120	1081	165	14.73%	96.52%	X
HSBC NORTH AMERICA HOLDINGS INC.	2649	2587	572	21.59%	97.66%	X
Seterus, Inc.	1436	1429	328	22.84%	99.51%	X

Initially the plan was to make a map which showed the over all data of all the companies on the map according to the state. Example Utah has 500 companies in banking sector and has total of 20000 complaints. But then we discussed and realized that the overall data has no use for the customer who want to see the performance of the company and not the complaints of the state. So what we thought of doing was on selection of the particular company from the bar chart its performance over different states were shown.

This might be useful of someone who keeps on travelling across different states and needs that company in that states. If the company is not present in the particular state then that state is shown by red color. If the company is doing bad in particular state then it is shown in dark shade compared to the other states where it is doing good. So basically we decided to implement a heat map over the states.

But the major question was “Why would most of the people care about the companies performance in different states if they don’t travel much and use the banking facilities mostly from one state?” This question can be answered by just linking the data both ways. First, by clicking the particular company in the bar chart its performance is shown in different states which is useful for someone who is travelling frequently whereas on clicking the particular state the top and bottom 10 companies of that states should be populated back in the bar chart.



Other major implementation change was we planned to use the stacked rectangles for each state (as shown above), where the entire rectangle showed the total complaints and smaller rectangle inside would show the selected criteria like disputed complaints and the timely

responded complaints. But this would look very clumsy. Also, the consumer might not be interested in just the total complaints and would be interested in the percentage of the disputed and the timely responded complaints with respect to the total complaints.

So, Shlok implemented a heat map which looks simple and still gives all the information one is looking for. Also there is tooltip on each state which shows all the information like State Name, total complaints, percentage of the disputed complaints over the total complaints and the percentage of the timely solved complaints over the total complaints. If the company is not present in the particular state then it will show “company not present in that state” and will be highlighted by the red color to make it clearly distinguishable as shown below

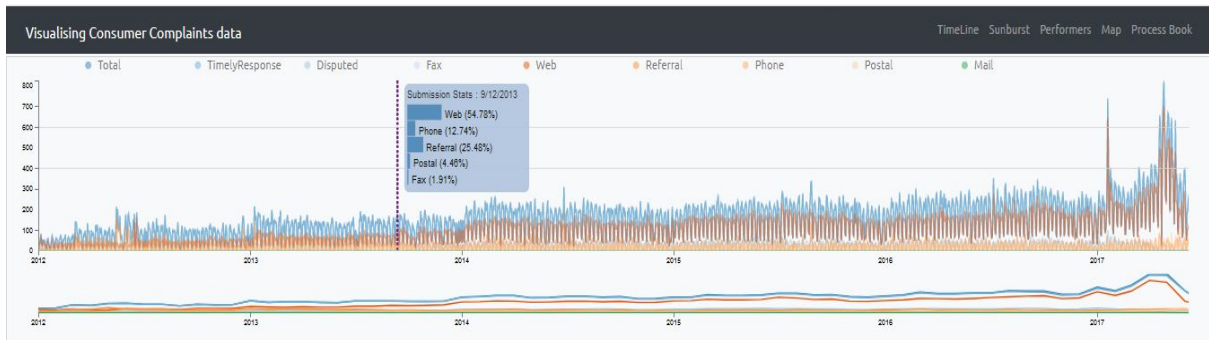


Nov 28, 2017: Final Project Team Meeting

We met for the last time to review the code, integrate it and test it as a single unit. We came up with minor upgrades to the visualization to improve the aesthetics and user interaction. The visualization served the high level purpose, but there were still a few minor bugs which needed to be taken care of.

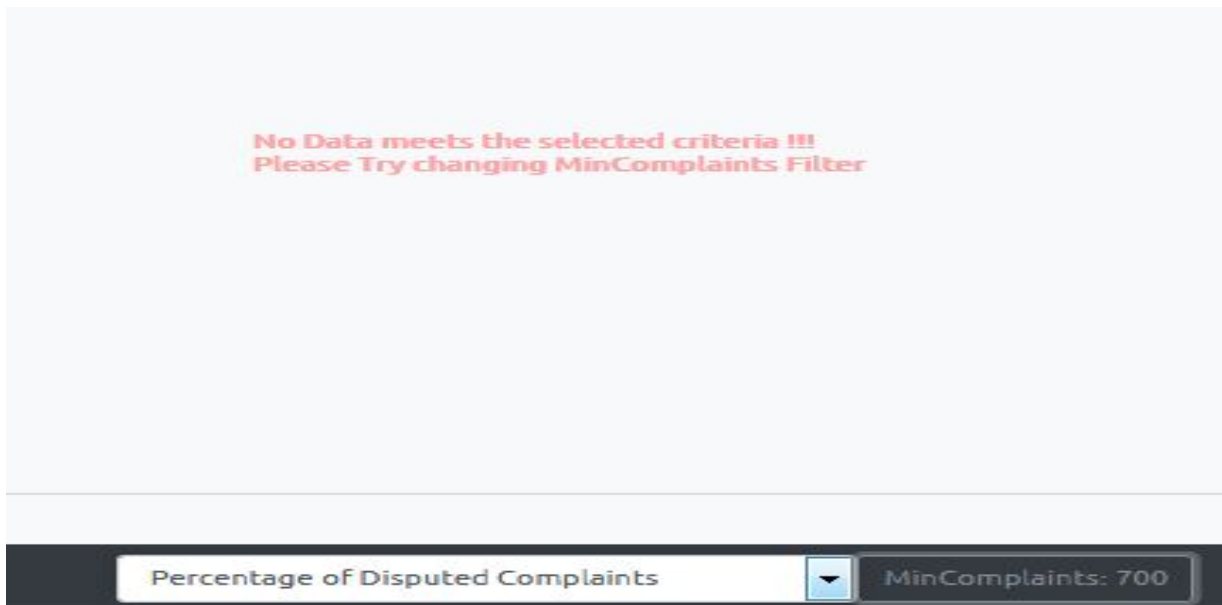
After resolving all the bugs, we decided to add a few last minute features to our visualization. This included having a tooltip and legends over the messy looking Timeline Chart, changing one of the view of the timeline chart to show data aggregated by month.

Initially, we designed the tooltip to show the statistics of the medium via which the complains were received. However, the change in the statistics could not be easily spotted. So, Madhur implemented a bar chart in the tooltip. This makes the change over time on mouse movement clearly visible. Since it is hard to spot a time series, we implemented a hover event which highlights the timeseries related to hovered legend item. Subsequently, the timeline chart looks as below :

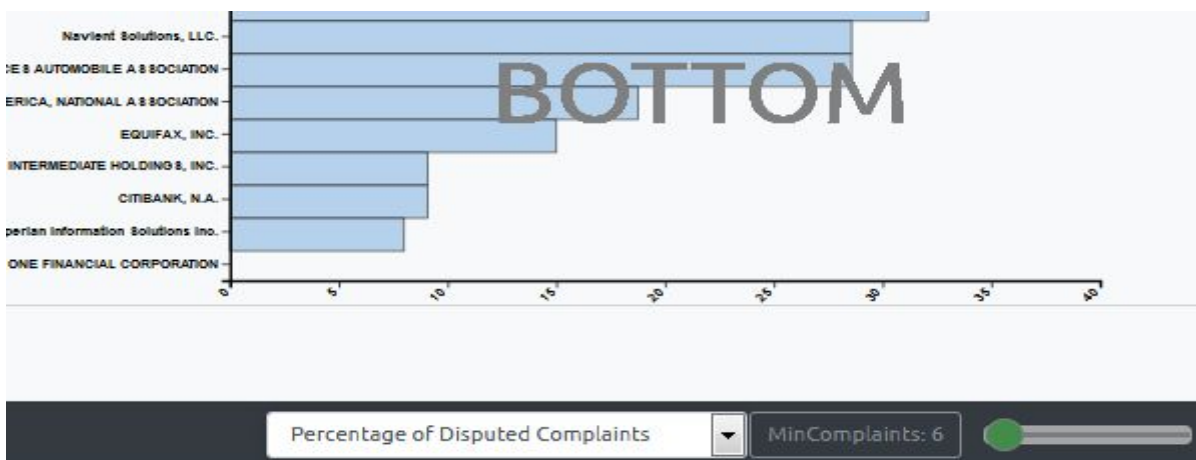


Also, in some of the cases there were no performers data populated after applying filter. This is because we stated a default value of 700 minimum complaints to evaluate a companies performance. To handle this scenario, we implemented a toggle button enabled slider in our control bar, which lets user change the minimum complaints criteria. This implementation is shown below

Before:

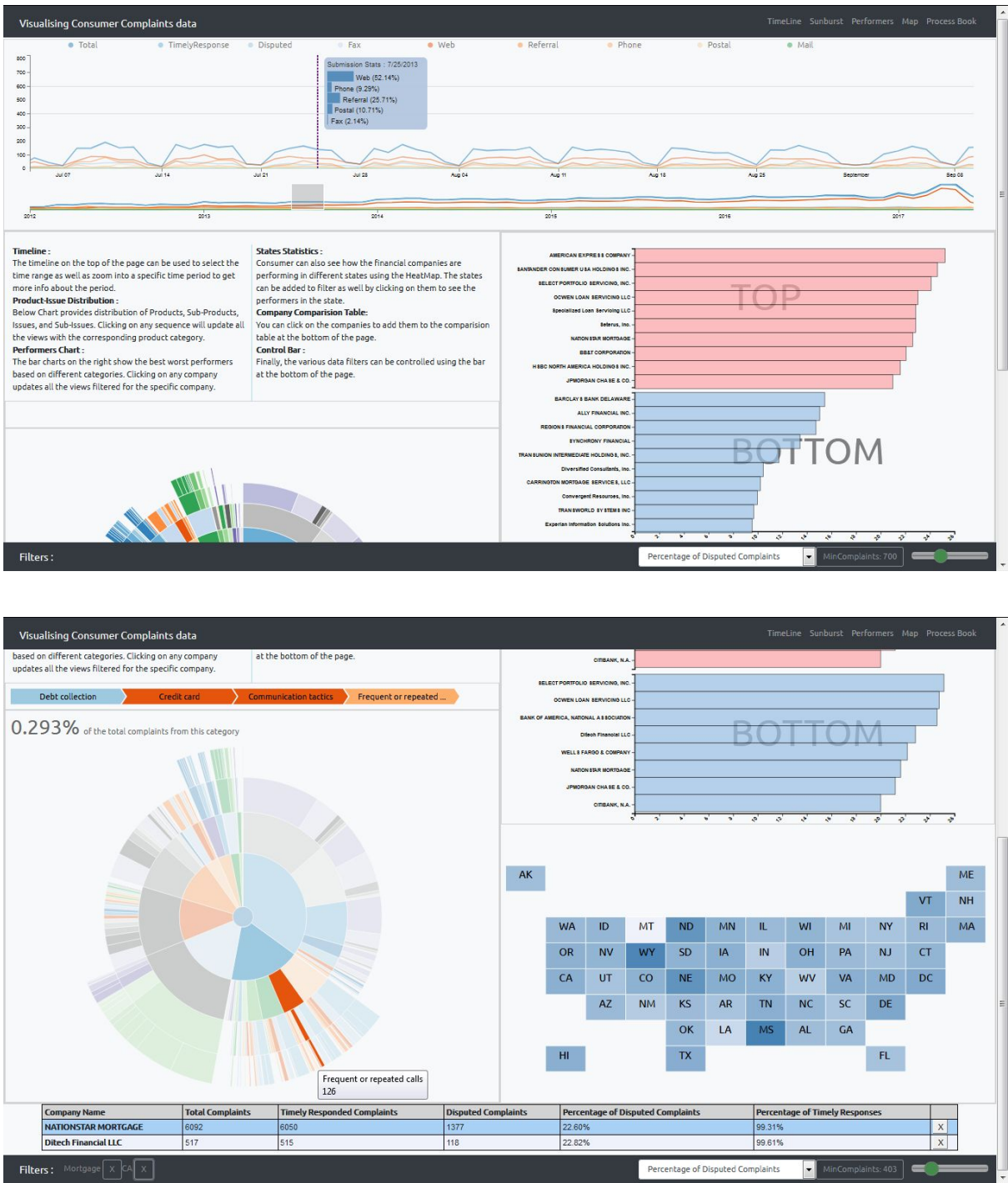


After:



Final Website:

Below is the snapshot of our entire website:



Evaluation:

As we worked towards our project, we found that having data in the correct form is the most important for any Visualization. We also believe that we could have done something better if

we had more data attributes rather than more records. This includes attributes like companies user base count and popularity index.

We were still able to implement the must-have features and the optional features that we thought would make our visualization effective.

From this visualization, we were able to see how a particular company is performing with it various product, in various states over the USA.

It also helps the end user looking for a specific product/service in finding out which service provider are good, and whom to be wary of, in his/her state. We are also able to see the trend of how people submit their complaints over time and how satisfied they are with the resolution.

It helps the companies management/regulators check which services are the customers most disappointed with and come up with strategies that could improve the specific product/service in the future. It also helps them compare their service against their top competitors.

Future Improvements:

The current implementation does not allow the end user to search for a specific company. In the future, we can have a search box which allows user to directly lookup the company of their choice and highlight the views accordingly. We could also improve the timeline chart to add/remove a series for individual companies.

Acknowledgements:

We are very thankful to <https://bl.ocks.org/>

The various visualizations here came in very handy in making the design choices and helped a lot during implementation.