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:

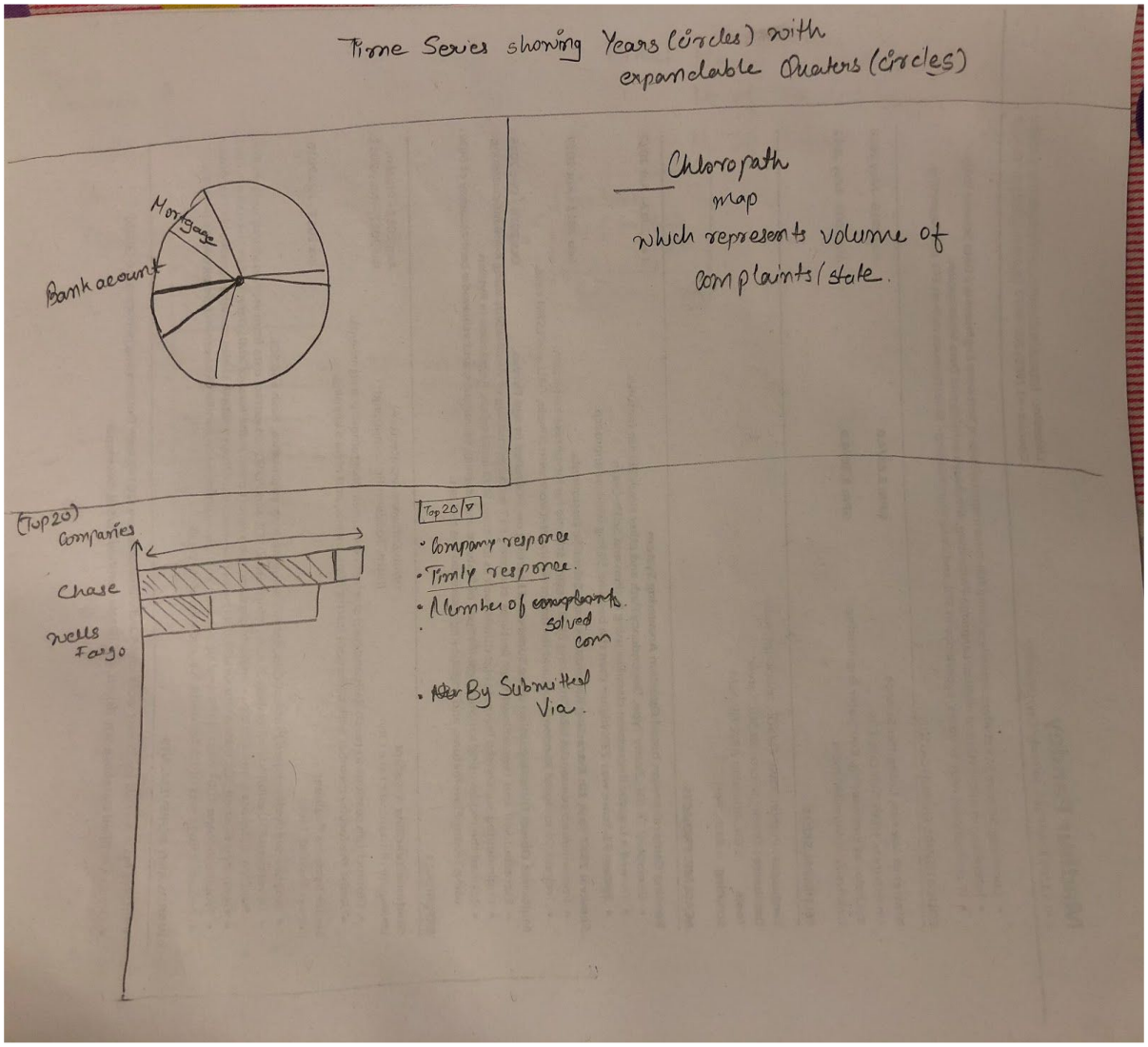
We might need minimal data processing. As of now, we only need to derive new date columns from the existing ones.
The dataset is very huge and we might have to select a random subset of the data so that it is faster in processing and represents the entire data. However, this is tentative.

Visualization Design:

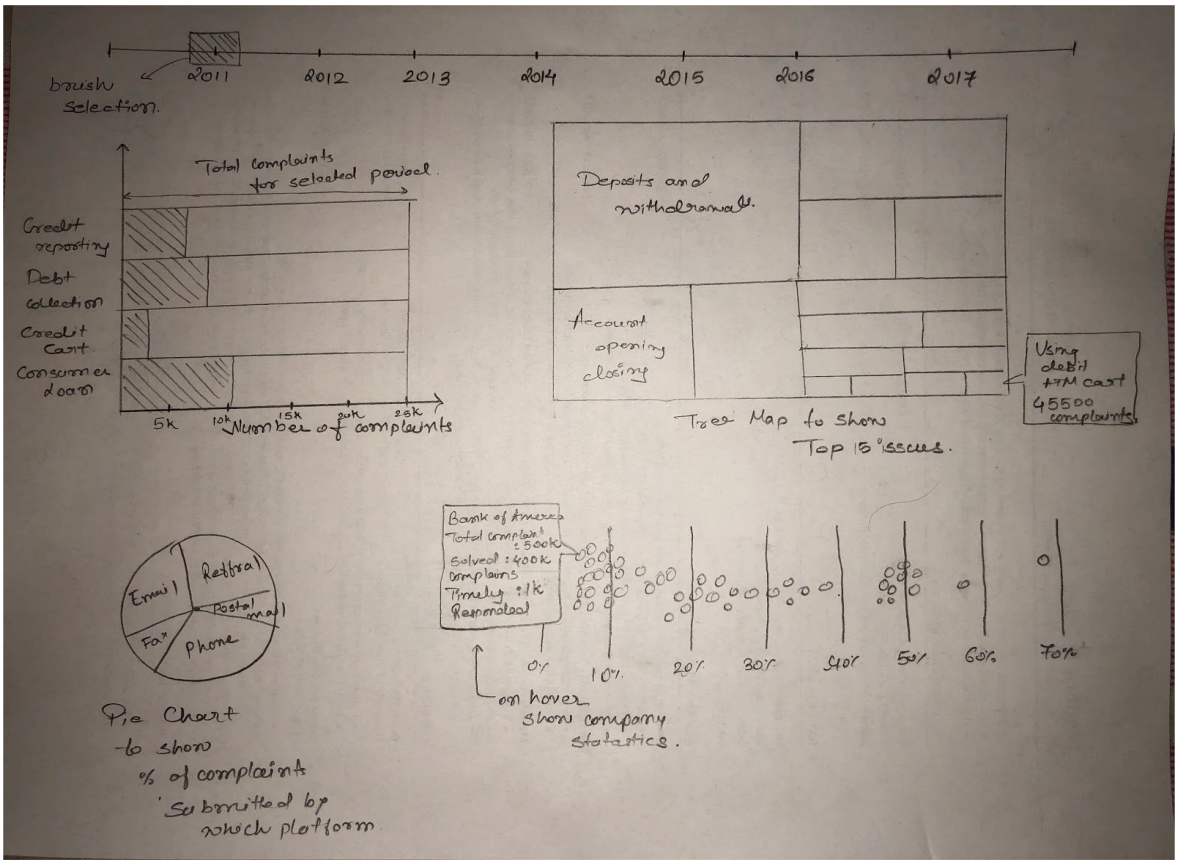
Our visualization design started with brainstorming the various questions we want our visualization to answer. After finalizing the questions, we came up with various initial designs and finally agreed upon a final design. Below are the snapshots of our initial designs and final design.

Initial Designs:

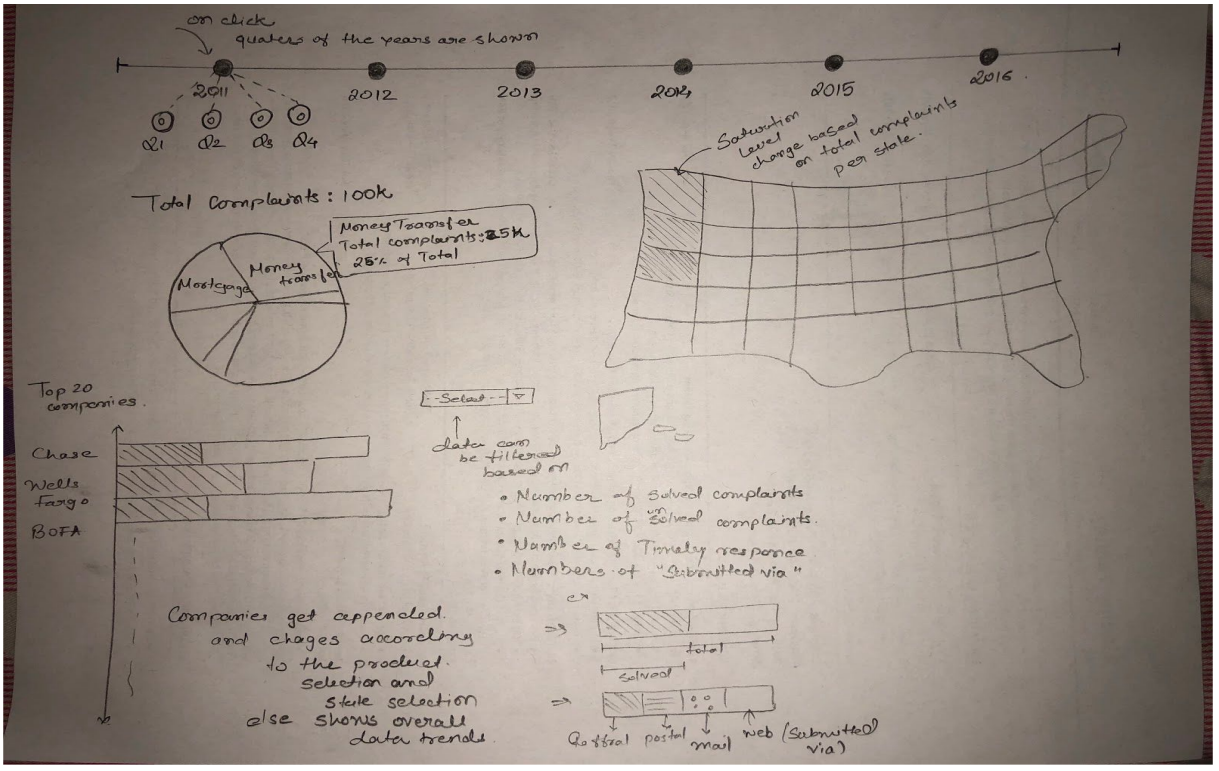
1.



2.


brush selection.
2011
2012
2013
2014
2015
2016
2017
Total complaints for selected period.
Credit reporting
Debt collection
Credit Card
Consumer loans
5k
10k
15k
20k
25k
Number of complaints
Deposits and withdrawals.
Account opening closing
Using debit ATM card
45500 complaints
Tree Map to show Top 15 issues.
Email
Referral
Postal mail
Fax
Phone
Pie Chart
-to show
% of complaints
submitted by
which platform
Bank of America
Total complaints : 500k
Solved : 400k
complains
Timely : 1k
Responsed
0%
10%
20%
30%
40%
50%
60%
70%
on hover
show company statistics.

3:


on click
quaters of the years are shown
2011
2012
2013
2014
2015
2016
Q1
Q2
Q3
Q4
Total Complaints : 100k
Money Transfer
Total complaints : 25k
25% of Total
Mortgage
Money transfer
Saturation level change based on total complaints per state.
Top 20 companies.
Chase
Wells Fargo
BOFA
Companies get appended and chages according to the product selection and state selection else shows overall later trends
Select
later can be filtered based on
Number of Solved complaints
Number of Solved complaints.
Number of Timely response
Numbers of "Submitted via"
total
solved
Referral
postal mail
web (Submitted via)

Final Design:

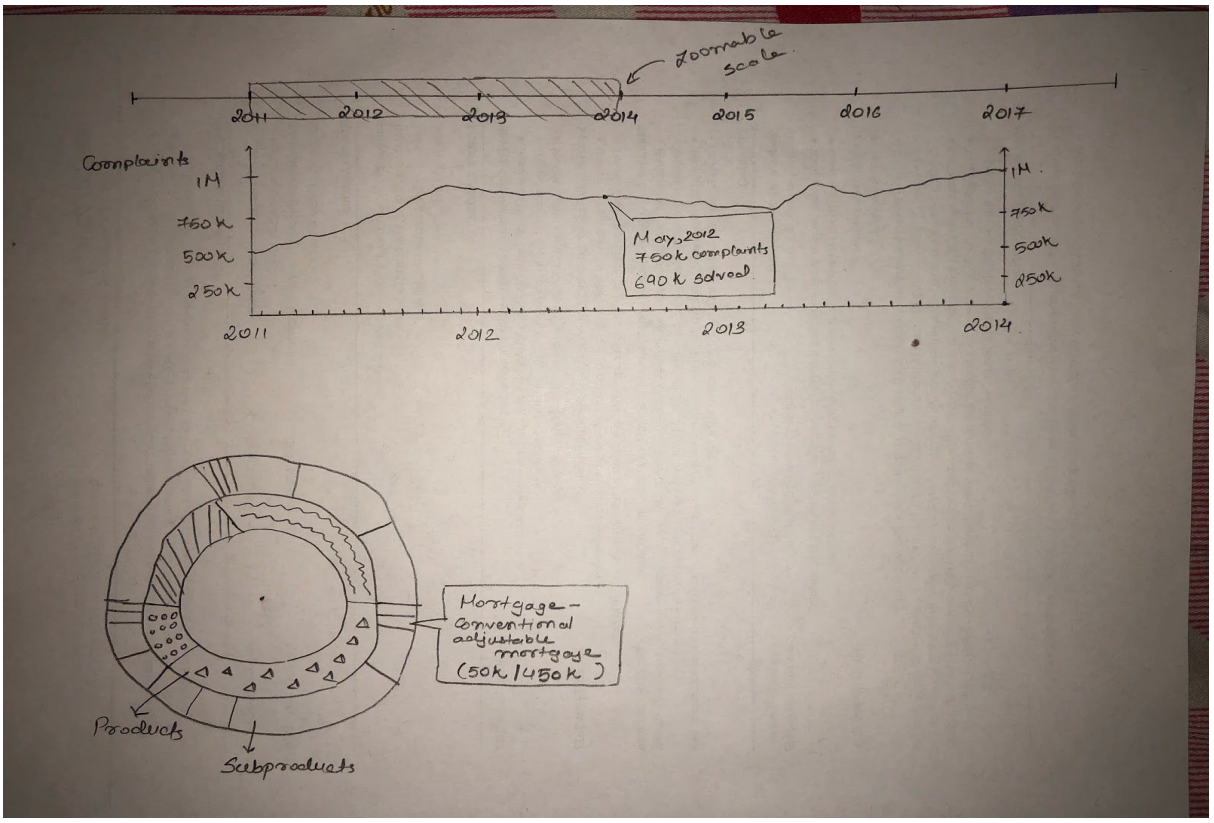


Fig 1: Complaints Trend over year and Product/Subproduct Sunburst Chart

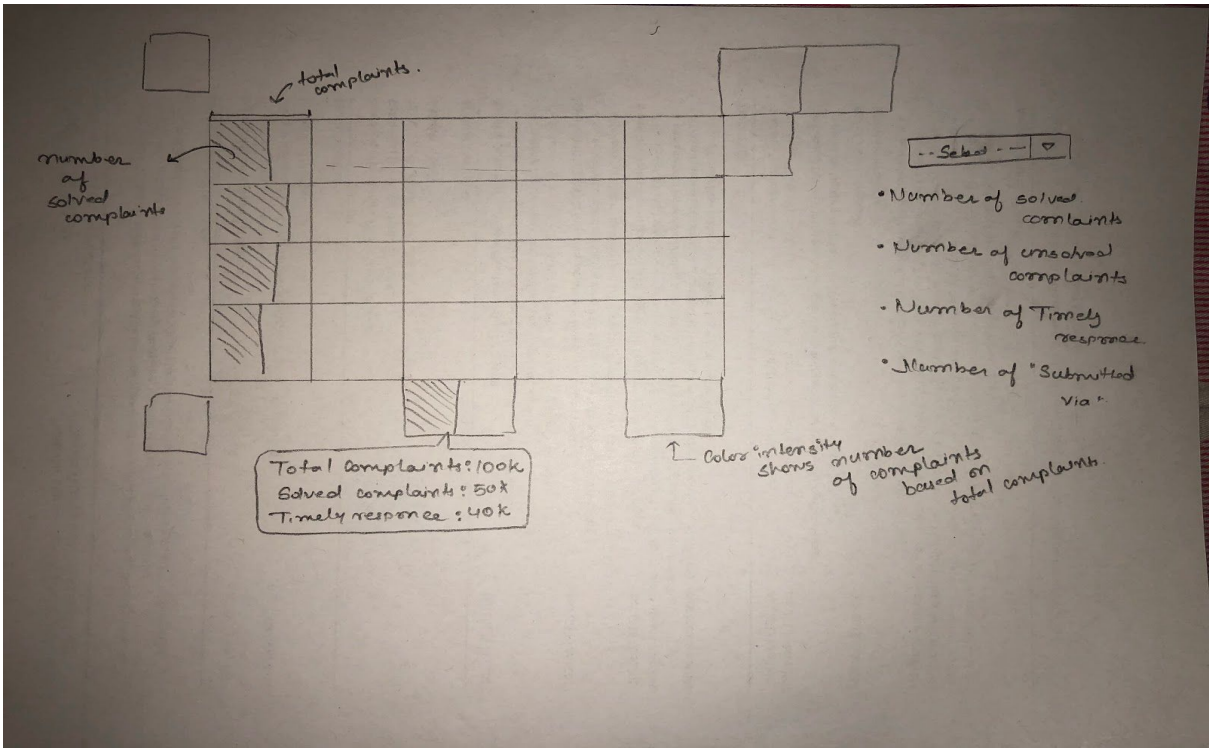


Fig 2: Map with area charts representing the overall state trend for the dropdown selected.

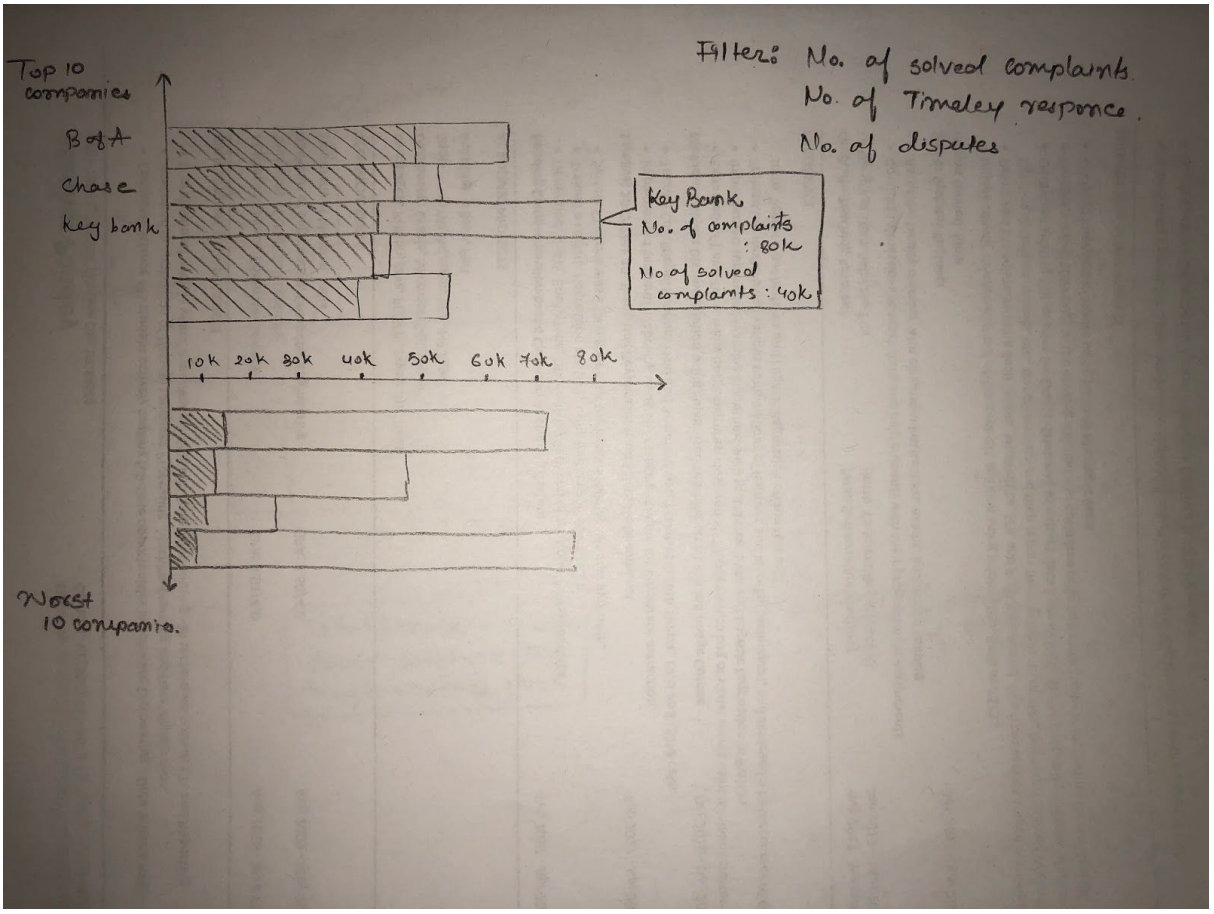


Fig 3: Stacked bar chart showing best/worst companies based on the dropdown value selected.

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 and 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:

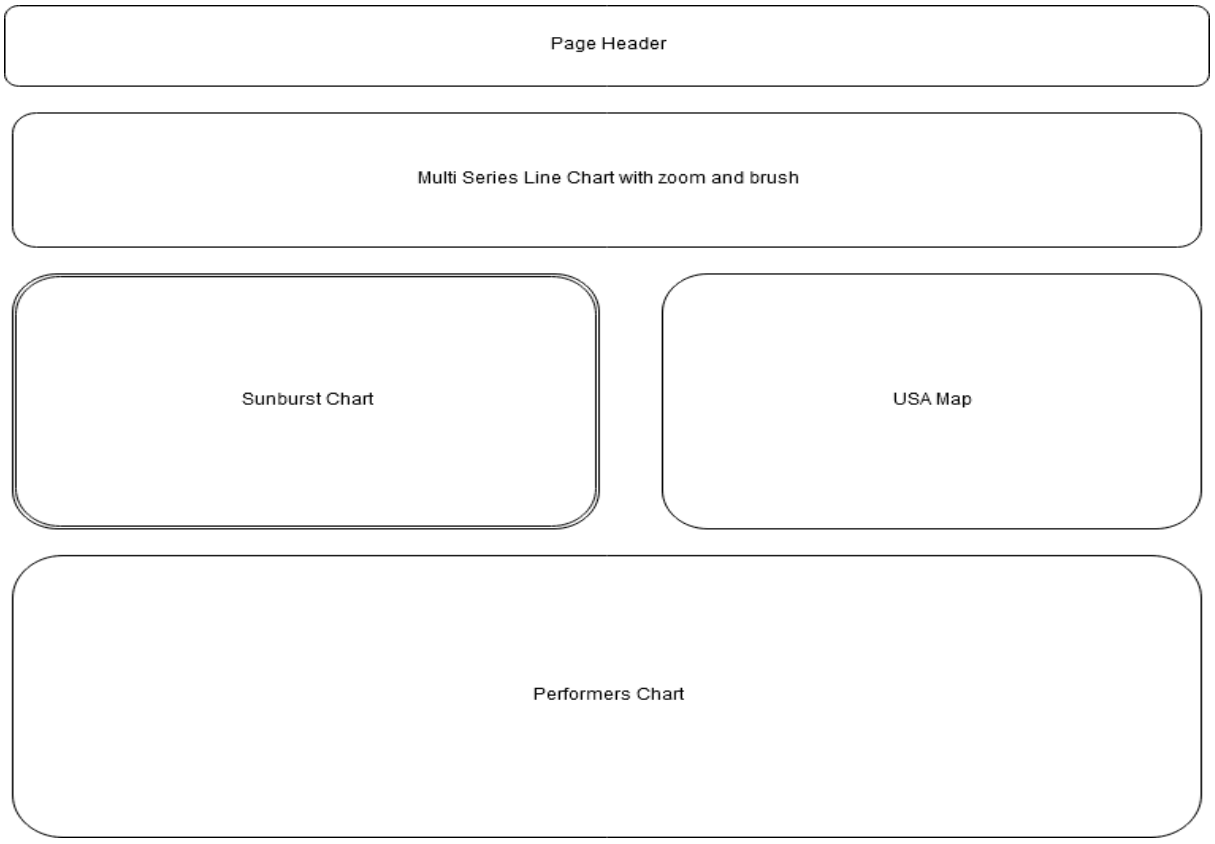
1. The overview visualization doesn't seem to be specific to the target audience and improving the same might help in getting better insights from the visualization.
2. Having a color scale in map (heat map) instead of bar chart, as bar chart might make the map look messy.
3. Customer might want to check 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

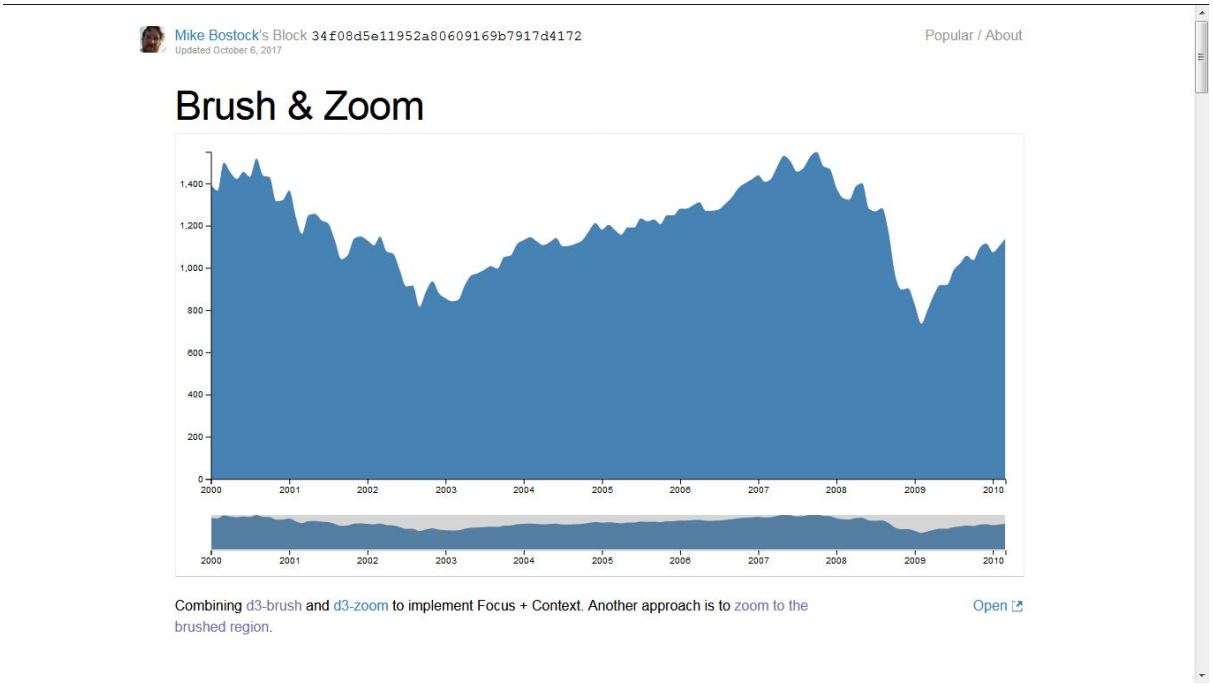
Both the team members meet up to further decide the approach to be used towards the implementation and check for any progress from any team members. Until now none of the member was able to put in dedicated hours towards the project.
It was decided in this meeting how the structure of the html file would be. We came up with the below structure for our webpage.



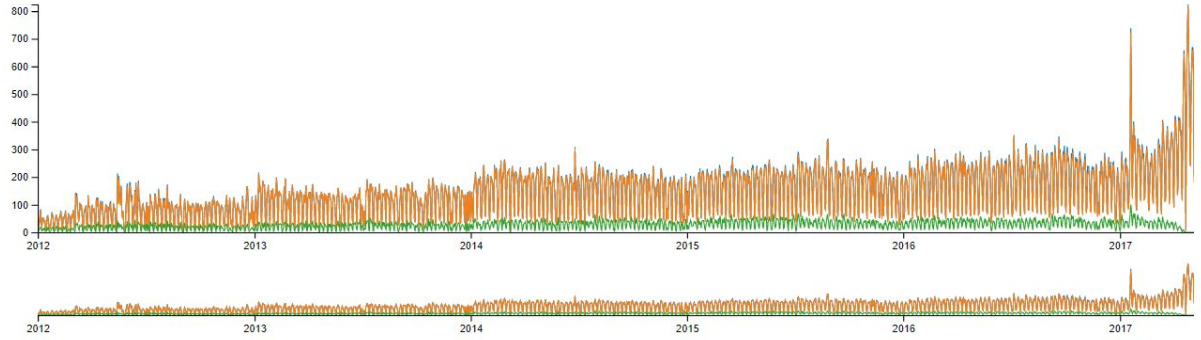
We then tried analyzing the data using excel. However, due to the large size (~400MB) of the csv file, we were unable to make much progress. We then turned towards Linux to truncate the file before using, however, due to newlines in the file, the truncation resulted in a file which was no longer in a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



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.

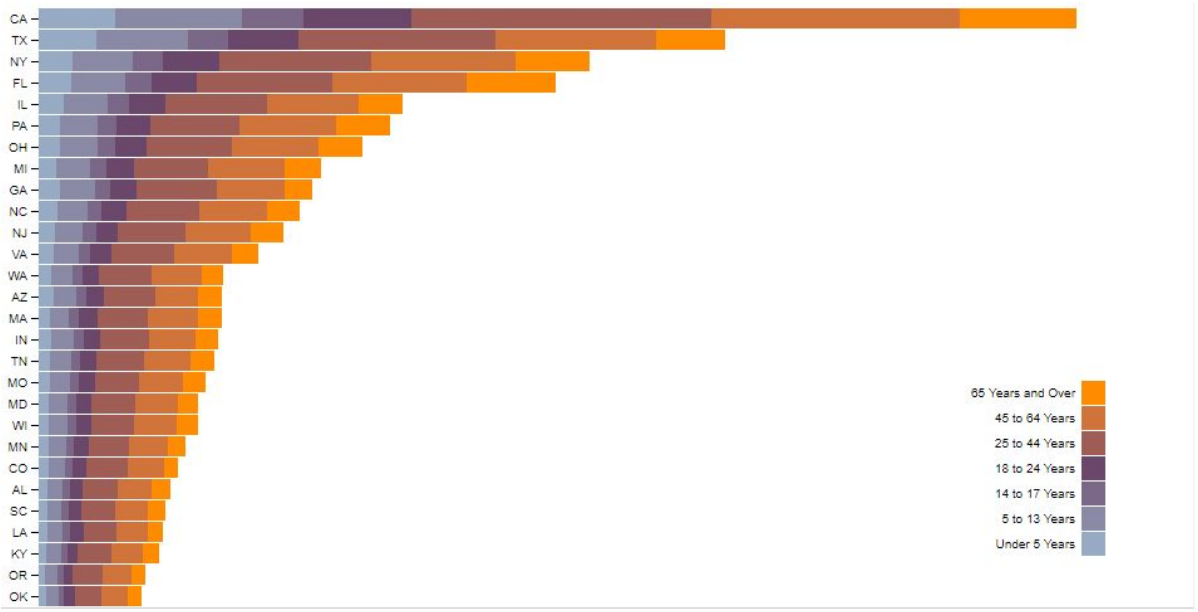
As of now, for different charts we are creating a custom data structure (JSON) which includes all the fields necessary for the making the visualization. Later on we will pass the corresponding data to other charts to make it interactive.

Nov 10, 2017:

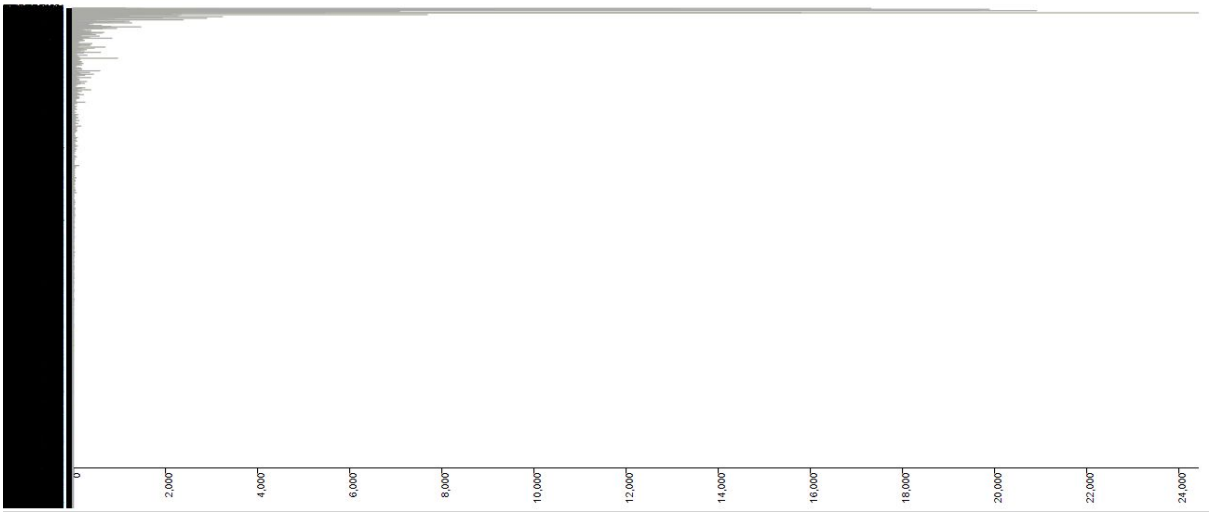
For the PerformanceChart.js, Shlok created the JSON object using nest() which includes "Company" field as key and "Timely Response", "Disputed Field”, “Submitted Via" and we are also calculating the total number of complaints for particular company and including it in the data structure as a "Total"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

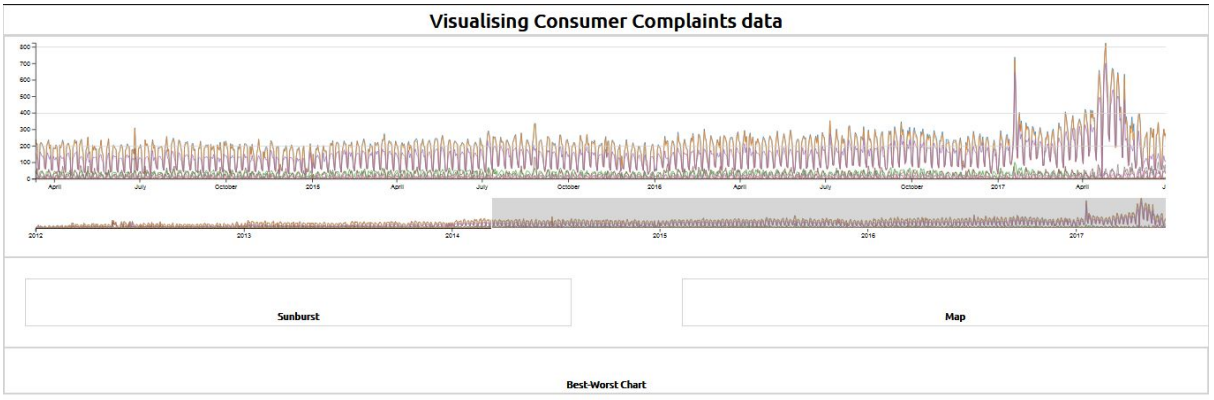


We are still working on it and its is near completion.We are currently getting 300k data over the y axis which should be sorted and we should filter only the top 10 and the worst 10 companies instead of all the companies. The initial implementation is shown below.



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 are the inner rectangle will be make 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.