**Data 608 - Knowledge and Visual Analytics**

**Final Project Proposal**

**Dataset:** Consumer complaint database.

**Brief explanation of the dataset:** This is a database of consumer complaints across financial products and services. The publisher of this data is Consumer Financial Protection Bureau. The variables available in the data are as follows:

* Date received – The date the complaint was received
* Product – The product type against which the complaint was raised
* Sub-product –The sub-product against which the complaint was raised
* Issue – What the complaint (Issue) was about.
* Sub-issue – Further drill down on the complaint.
* Consumer complaint narrative – Description of the complaint
* Company public response – What was the response from the company
* Company – The financial company
* State – Which state the complaint was raised
* ZIP code -The zip code
* Tags – Tags used to classify the complaint
* Consumer consent provided?
* Submitted via -What was the channel used to raise the complaint
* Date sent to company – The date when the complaint was sent to the company by CFPB
* Company response to consumer – Action taken by the financial company
* Timely response? – Was the response in a timely manner
* Consumer disputed? – Did the consumer dispute the response?
* Complaint ID – The complaint ID.

**Dataset characteristics:** This is a csv data that is downloaded from the below given URL. The dataset contains 742,451 records and the size of the dataset is close to 200 MB.

**Dataset URL:** <https://catalog.data.gov/dataset/consumer-complaint-database>

**Dataset processing:** A major challenge with the data is the huge size. Given that this data needs to be rendered on the web (A shiny App) and that the data would most probably reside in my github account, the challenge would be to resolve the graphics in a timely manner.

To limit the size of the data, I have decided to preprocess this dataset and consider only the following variables for the final analysis and condense the dataset based on the count of records for the below variable combinations:

* Date received – The date the complaint was received
* Product – The product type against which the complaint was raised
* Issue – What the complaint (Issue) was about.
* Company public response – What was the response from the company
* Company – The financial company
* State – Which state the complaint was raised
* Submitted via -What was the channel used to raise the complaint
* Company response to consumer – Action taken by the financial company
* Timely response? – Was the response in a timely manner
* Consumer disputed? – Did the consumer dispute the response?

Some of the preprocessing would involve the following:

1. Bin the data into years and months and taking counts (further reducing the data size)
2. Keep only relevant products and company information and the counts there of (any product or company with less numbers would be classified into “others”)

In addition for each analysis, I would like to create a separate condensed dataset catering only to that analysis.

**Few analyses I have in mind:** The following are a few analysis / charts I have in mind as of now:

1. Trend of complaints across years by state, product, etc. – Here, I would like to show a time-lapse chart comparing either the states or products or submission channels
2. Which states have the most complaints filtered by product, company and year?
3. Has the company improved over the years?
4. What channels are used to complain across products, company or in a state?
5. How timely is the company response to complaints?
6. Which companies did not resolve the issue satisfactorily (based on the consumer disputing the solution)

I am sure that I will be able to add a few as I progress on the project

**Relevance of the Analysis:** This project is about data exploration and to find trends and patterns in the data that is available to us. Though the analysis will be at a high level, it will enable the companies to identify areas of focus and improvement in their customer service and conflict resolution. This will lead to a more satisfied customer.

**Technologies to be used:** I amplanning to use Shiny as the primary driver of the analysis. Each “tab” in shiny would address a particular area of analysis. Each analysis in turn will use either plotly, ggplot, etc. based on the type of chart and interaction I wish to provide in the chart. Github will be used to store all the data assets. I would strive to use all technologies that we have covered in the course.