CAPSTONE Project: Improving and Gathering Insights from Consumer Complaints from Consumer Federal Protection Bureau

**Problem Statement:**

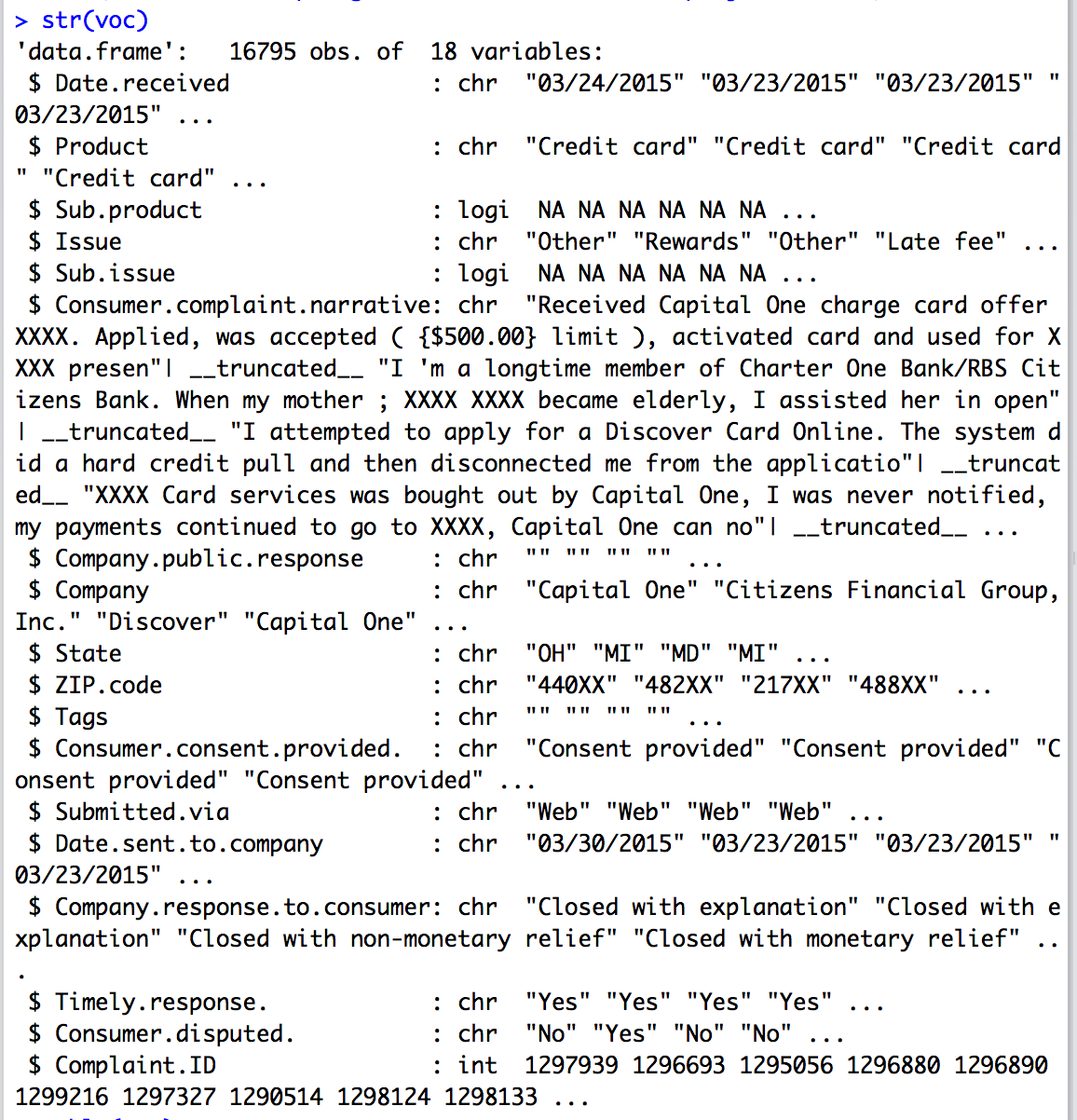
Various financial industries are affected by higher consumer confidence index which measures how optimistic U.S. consumers are at a specific point in time. While economists and investors can debate just how significant this indicator is, most at least grudgingly agree that how consumers feel about the economy (and their personal financial situation can be a self-fulfilling prophecy.

CFPB states, “by submitting a complaint, consumers can be heard by financial companies, get help with their own issues, and help others avoid similar ones.”(CFPB website) Every complaint provides insight into problems that people are experiencing, helping us identify inappropriate practices and allowing us to stop them before they become major issues. The result: better outcomes for consumers and a better financial marketplace for everyone.

Data Set: Consumer Complaint Database API DOCS:

Each week CFPB send thousands of consumers’ complaints about financial products and services to companies for response. Complaints are listed in the database after the company responds or after they’ve had the complaint for 15.

Important information in the data set are the text complaints (consumer complaint narrative)



Data Wrangling:

Once the datasets are imported, the following data cleanup activity is performed –

1. Consumer Complaint Narrative: The CC narrative column is imported as a character variable and has XXX symbol representing people account credit card numbers or account. The xxx value is removed using gsub & the column is converted to a character factor. I had to get rid of any punctuations, spaces, dollar signs and slashes so that the text can be converted into a corpus.
2. Company public response: I have to get rid of the space between the text. Also I have to get rid of an punctuations.
3. Timely Response and Time Received: Both variables need to be converted is set to character. It is converted to numeric type and plot to see the relationship
4. Handling NA values All the NA values in the data are set to the mean of the corresponding columns  The next steps will be to build the clustering models and discover the new customer groups on the testing dataset.
5. Zip Code: The variable describes the where the consumer has filed the complaint from what area and zip code. However, due to privacy, CFPB has used zip codes the first three.