A red and black logo

Description automatically generatedGeneral Assembly

Data Analytics Immersive Bootcamp

Consumer Financial Protection Bureau

**Qasim Malalla**

**3/7/2025**

# 1.0 Introduction

Every complaint in the CFPB database represents a consumer who felt misrepresented or mistreated by the financial system. While institutions may see tickets and case numbers, this project views complaints as a story—a reflection of systemic gaps and unmet expectations. With over 500,000 complaints from 2020 to 2024, this analysis aims to understand and amplify these voices to identify trends, gaps, and opportunities for action.

# 1.1 Problem Statement

Despite the large volume of complaints submitted to the CFPB, these data are underutilized in uncovering systemic weaknesses in financial services. The lack of timely, actionable analysis results in missed opportunities for policy reform, institutional improvement, and consumer empowerment.

# 1.2 Objectives

- Identify complaint trends by time (month, weekday, year) to uncover behavioral or systemic triggers.

- Analyze which financial products and issues generate the most consumer frustration.

- Benchmark companies based on complaint volume, timeliness of response, and resolution success.

- Derive actionable insights to inform regulators, improve company response strategies, and protect consumers.

# 1.3 Target Audience

- Regulators (e.g., CFPB, financial oversight bodies)

- Financial Institutions and banks

- Consumers and consumer rights advocates

- Researchers in finance, public policy, and data science

# 1.4 Dataset(s)

- Source: Consumer Financial Protection Bureau (CFPB)

- Period Covered: 2020–2024

- Size: 500,000 complaints, 2,904 unique companies

- Geographic Scope: United States

- Submission Method: Web

- Data Dictionary Highlights:

- Date received, Product, Sub-product, Issue, Narrative, Company, State, Submitted via, Timely response?

# 1.5 Data Handling

- Cleaning: Removed nulls and inconsistencies; standardized categorical fields (e.g., product names).

- Transformation: Extracted date parts (month, weekday, year); grouped company variants.

- Handling Missing Values: Focused narrative analysis only on ~174,000 non-null entries.

- Preprocessing for NLP: Cleaned text (lowercasing, stopwords removal, punctuation stripping).

- Embedding & Clustering: Used `all-MiniLM-L6-v2` model and clustered narratives for topic discovery.

# 1.6 Analysis and Findings

- Temporal Trends: Spike in complaints during Q4 (Oct–Dec).

- Geographic Distribution: Complaint volume correlates with population and economic activity.

- Top Issues: Credit reporting dominates complaints.

- Company Performance: Credit bureaus have highest complaint volumes; varied resolution success.

- Submission Behavior: Most complaints submitted online; users expect fast response.

# 1.7 Recommendations

- Operational Preparedness: Improve Q4 support and fraud systems.

- Transparency Tools: Create clearer dashboards for users to verify/report errors.

- Lifecycle Modernization: Build modern, AI-enhanced complaint management systems.

- National Intelligence Framework: Monitor and act on complaint trends nationwide.

# 1.8 Limitations and Assumptions

- Limitations:

- Only includes published complaints.

- May underrepresent certain demographics.

- Some complaints lack resolution data.

- Assumptions:

- Complaint count reflects real dissatisfaction.

- Timely responses align with better outcomes.

- Narrative text provides useful insights.

# 1.9 References

- Consumer Financial Protection Bureau Dataset: https://catalog.data.gov/dataset/consumer-complaint-database

- Sentence Transformers: https://www.sbert.net

- U.S. Census Bureau (for per capita insights)

- Equifax, Experian, TransUnion annual reports

# 1.10 Data Dictionary

|  |  |
| --- | --- |
| Column Name | Description |
| Date received | The date the CFPB received the consumer complaint. |
| Product | The financial product the complaint is about (e.g., credit card, mortgage). |
| Sub-product | More specific category under the product (e.g., checking account under bank account). |
| Issue | The type of problem the consumer experienced (e.g., incorrect information on report). |
| Sub-issue | A more detailed description of the issue. |
| Consumer complaint narrative | The optional free-text description written by the consumer. |
| Company public response | The response the company gave that was published by the CFPB. |
| Company | The name of the company the complaint is about. |
| State | The U.S. state where the consumer resides. |
| ZIP code | The ZIP code of the consumer. |
| Tags | Additional information (e.g., “Servicemember,” “Older American”). |
| Consumer consent provided? | Whether the consumer gave consent to publish their narrative. |
| Submitted via | How the consumer submitted the complaint (e.g., Web, Phone, Referral). |
| Date sent to company | The date the complaint was forwarded to the company. |
| Company response to consumer | The final response the company gave to the consumer. |
| Timely response? | Indicates whether the company gave a timely response. |
| Complaint ID | Unique identifier for each complaint. |
| Year\_received | Extracted year from Date received. |
| Month\_received | Extracted month from Date received. |
| Weekday\_received | Extracted weekday from Date received. |
| Year\_sent | Extracted year from Date sent to company. |
| Month\_sent | Extracted month from Date sent to company. |
| Weekday\_sent | Extracted weekday from Date sent to company. |
| CFPB to Company | Time taken (in days) for the CFPB to forward the complaint to the company. |
| has\_narrative | Binary flag (1/0) indicating whether a narrative is present. |
| clean\_text | Preprocessed version of the narrative (lowercase, no stopwords/punctuation). |
| narrative\_length | Word count of the cleaned narrative. |
| sentiment\_score | Sentiment polarity score from -1 (negative) to 1 (positive). |
| is\_fraud\_related | Flag indicating presence of fraud-related keywords in the narrative. |
| complaints\_per\_100k | Complaint rate per 100,000 people in a ZIP/state, used for normalization. |

# 1.11 Dashboard

A screenshot of a computer screen

Description automatically generated