Individual Final Project Report

NLP for Financial Consumer Protection

Name: Amogh Ramagiri
Group Members: Lasya Raghvendra, Vaijayanti Deshmukh

1. Introduction

Project Overview
Our group project aimed to automate the classification of consumer complaints for the Consumer Financial Protection Bureau (CFPB). We utilized Natural Language Processing (NLP) to route unstructured complaint narratives into specific product categories (e.g., "Mortgage," "Credit Card"), thereby reducing manual triage costs.

1.1 Outline of Shared Work

The group collaboratively performed the initial data acquisition, cleaning, and Exploratory Data Analysis (EDA). We worked together to define the scope, selecting the specific "Product" target variable and handling the significant class imbalance present in the dataset.
Outline of Individual Contribution

My specific contributions to this project are:

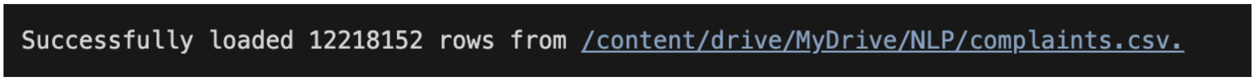
1. This report presents an in-depth Exploratory Data Analysis of the CFPB (Consumer Financial Protection Bureau) consumer complaints dataset.
2. The goal is to understand dataset structure, assess data quality, examine distributions, identify patterns across products and issues, analyze narrative text characteristics, and evaluate class imbalance to support downstream NLP modeling.

2. Data Loading and Structure

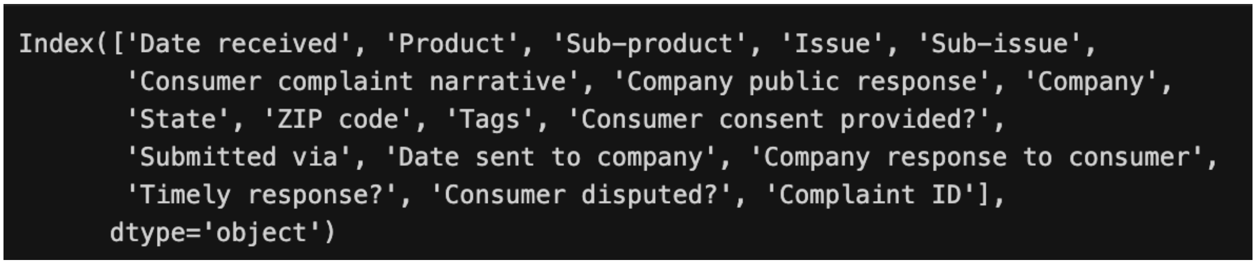
2.1 Columns Overview

Columns include complaint metadata (dates, company, state, etc), categorical labels (product, issue), consumer responses, and text narratives.

The “.columns” output lists columns such as: Date received, Product, Sub-Product, Issue, Subissue, Consumer complaint narrative, etc.



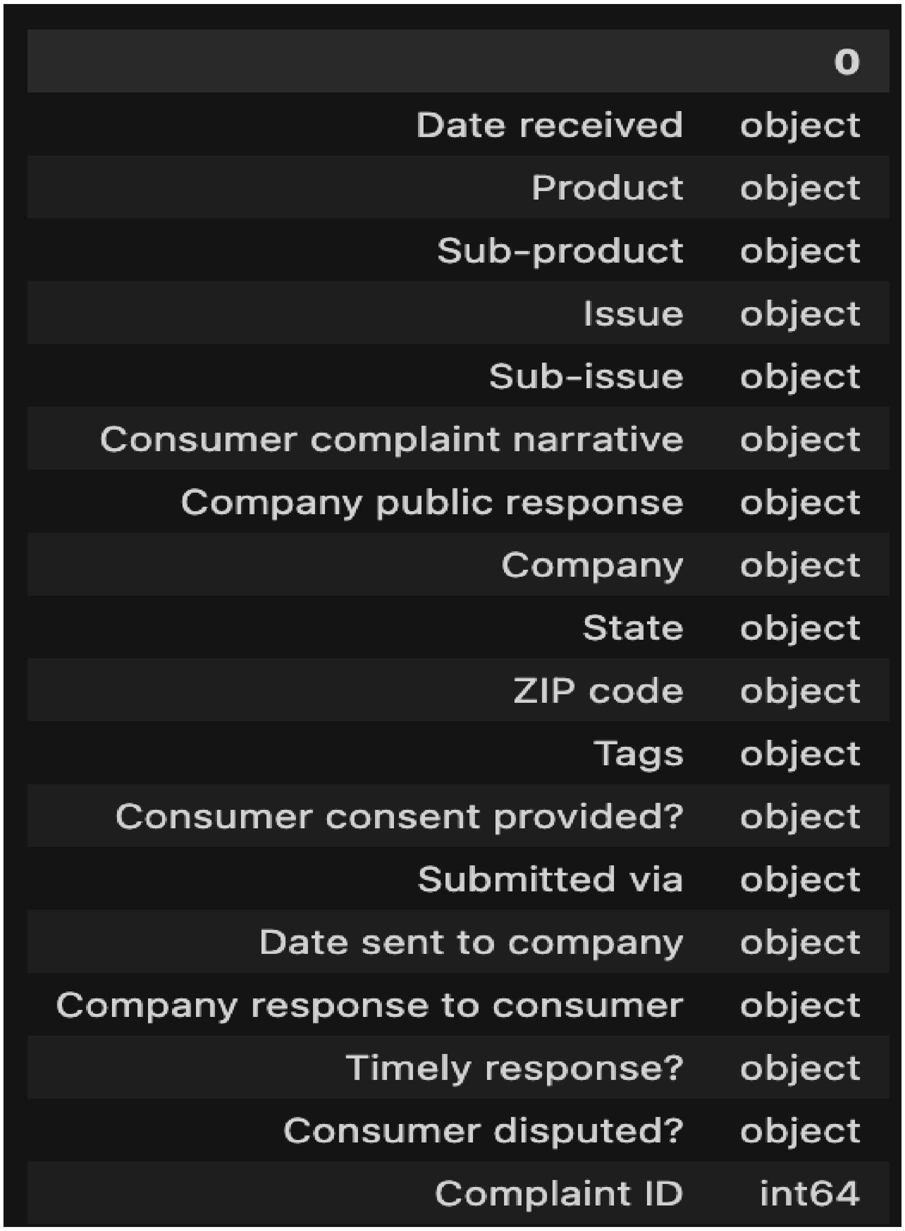
```
Successfully loaded 12218152 rows from /content/drive/MyDrive/NLP/complaints.csv.
```



```
Index(['Date received', 'Product', 'Sub-product', 'Issue', 'Sub-issue',
       'Consumer complaint narrative', 'Company public response', 'Company',
       'State', 'ZIP code', 'Tags', 'Consumer consent provided?',
       'Submitted via', 'Date sent to company', 'Company response to consumer',
       'Timely response?', 'Consumer disputed?', 'Complaint ID'],
      dtype='object')
```

2.2 Data Types

Most columns are of type object, except Complaint ID, which is int64.



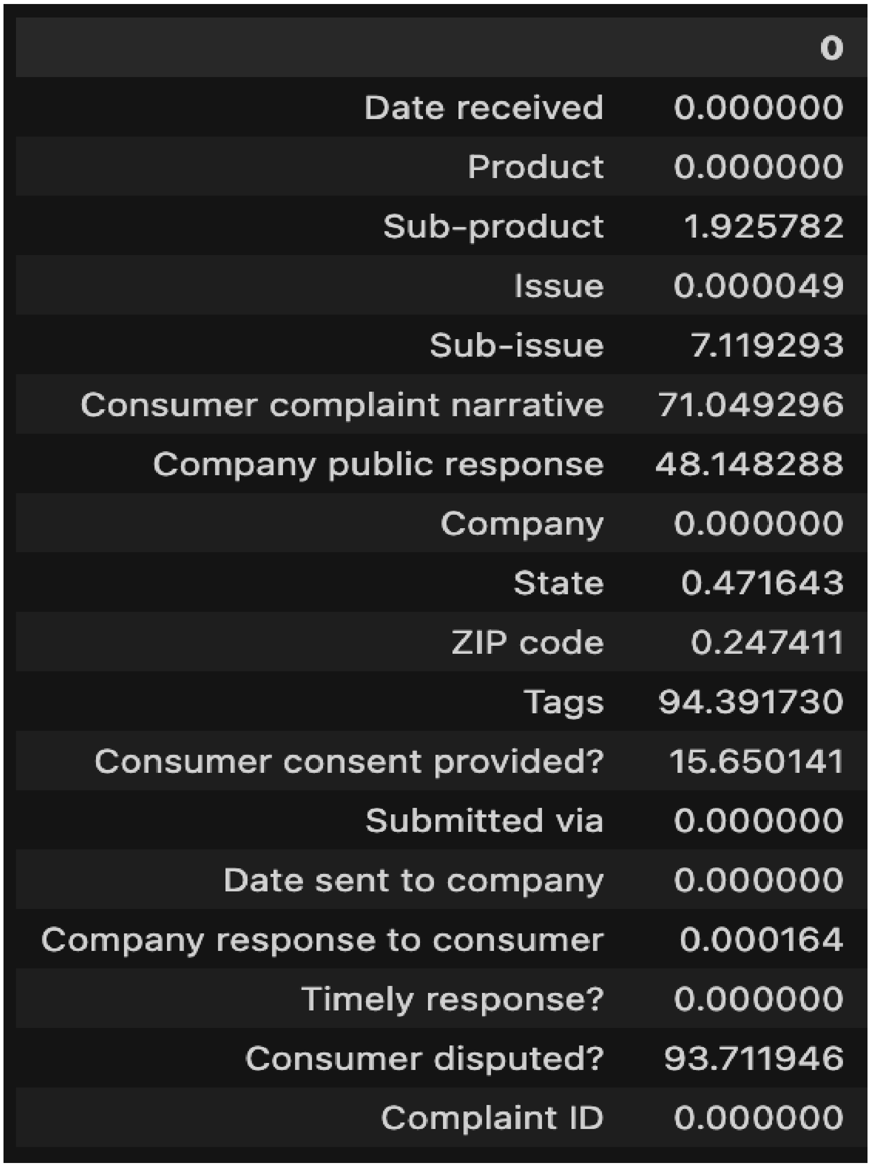
	0
Date received	object
Product	object
Sub-product	object
Issue	object
Sub-issue	object
Consumer complaint narrative	object
Company public response	object
Company	object
State	object
ZIP code	object
Tags	object
Consumer consent provided?	object
Submitted via	object
Date sent to company	object
Company response to consumer	object
Timely response?	object
Consumer disputed?	object
Complaint ID	int64

3. Data Quality Assessment

3.1 Missing Values

- Several columns contain significant missing values:
    1. **Consumer complaint narrative:** 71.05% missing (8.68million records)
    2. **Tags:** 94.39% missing
    3. **Company public response:** 48% missing
    4. **Consumer consent provided:** 15.65% missing

This reveals the need for filtering when working with textual narratives.



	0
Date received	0.000000
Product	0.000000
Sub-product	1.925782
Issue	0.000049
Sub-issue	7.119293
Consumer complaint narrative	71.049296
Company public response	48.148288
Company	0.000000
State	0.471643
ZIP code	0.247411
Tags	94.391730
Consumer consent provided?	15.650141
Submitted via	0.000000
Date sent to company	0.000000
Company response to consumer	0.000164
Timely response?	0.000000
Consumer disputed?	93.711946
Complaint ID	0.000000

3.2 Duplicate Records

No duplicate rows found (0 duplicates)

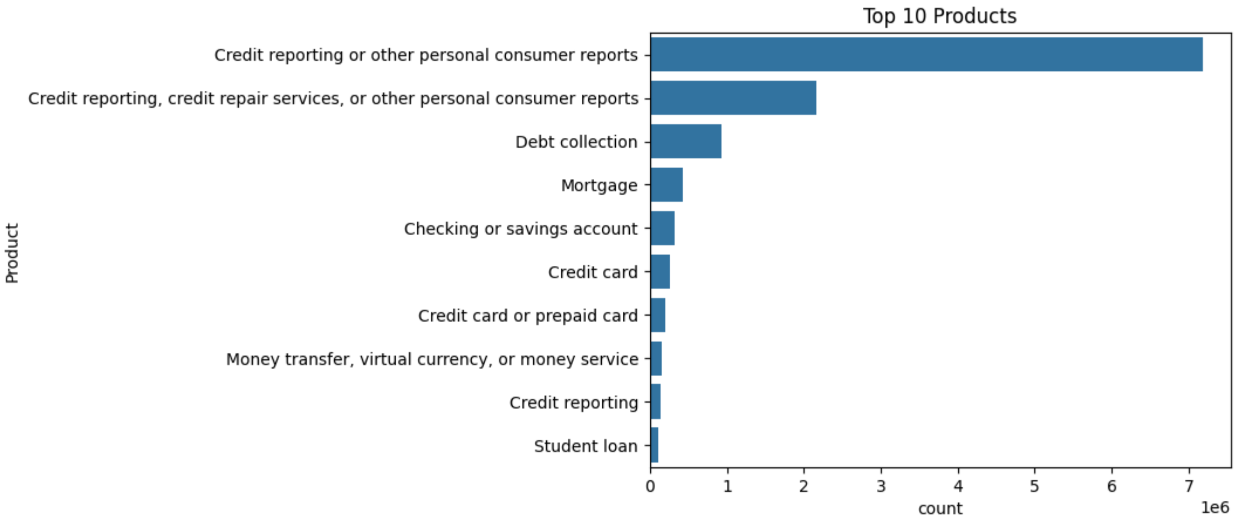
4. Univariate Analysis

4.1 Product Distribution

The top product categories are highly skewed:

1. **Credit reporting or other personal consumer reports** – 7,187,113
2. **Credit reporting, credit repair services, or other personal consumer reports** – 2,163,844
3. **Debt collection** – 929,544
4. **Mortgage** – 433,092

This imbalance is a major challenge for classification models.



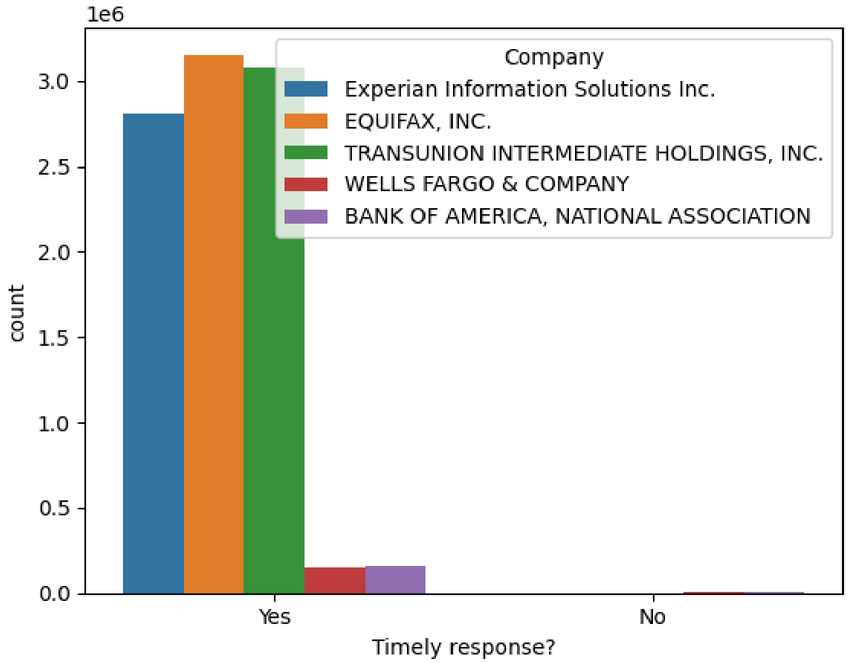
4.2 Issue Distribution

Issues are spread across Incorrect information, Account management, Struggling to pay debt, etc. Cross-tab output shows many issues linked specifically to cred related products.

4.3 Company Frequency

Top companies include:

- Experian
- Equifax
- TransUnion
- JPMorgan Chase
- American Express



4.4 Submission Method

Web dominates submissions with 11.6 million complaints, followed by:

- Referral (266k)
- Phone (208k)
- Postal mail (105k)

4.5 Boolean Columns

- **Timely response?** → 97% "Yes" (approx.)
- **Consumer disputed?** → 19.31% "Yes"

5. Bivariate Analysis

5.1 Product X Issue Cross-Tab

Attached output shows which issues are dominant within each product.
Example: Credit card complaints heavily include *advertising and marketing issues* (e.g., 6,168 complaints) .

5.2 Company X Timely Response

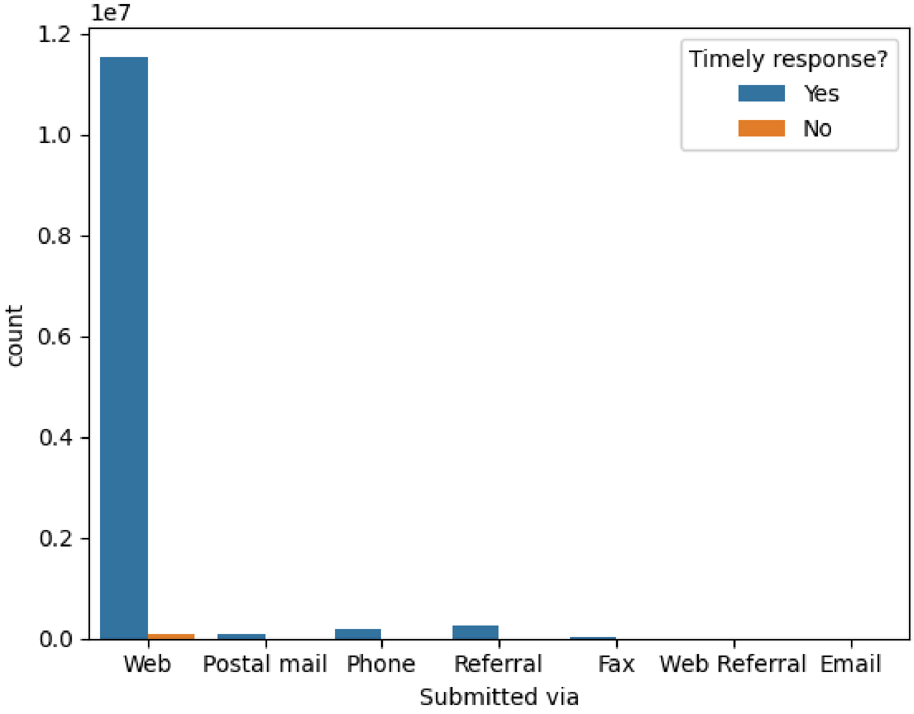
Visual output shows companies vary slightly in timely response rate, but most have high compliance.

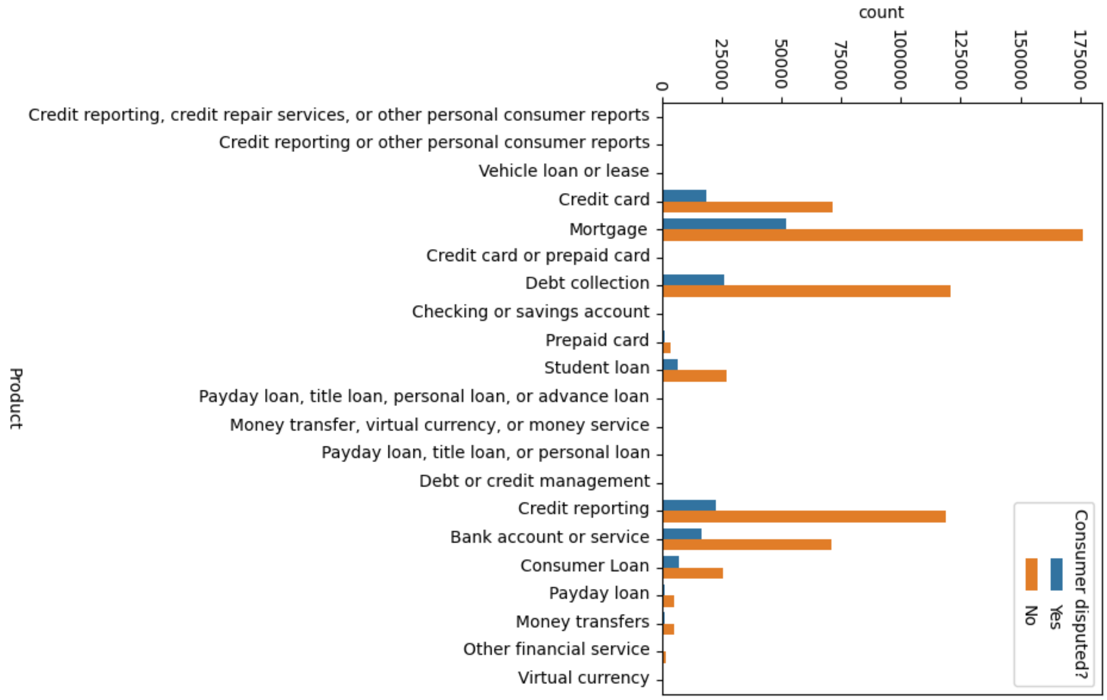
5.3 Submitted X Timely Response

Web and referral submissions receive responses faster than fax/postal submissions. Countplot supports this trend.

5.4 Product X Consumer Disputed?

Products like credit reporting have higher dispute rates than credit card or mortgage categories. Plot indicates non-uniform distribution.

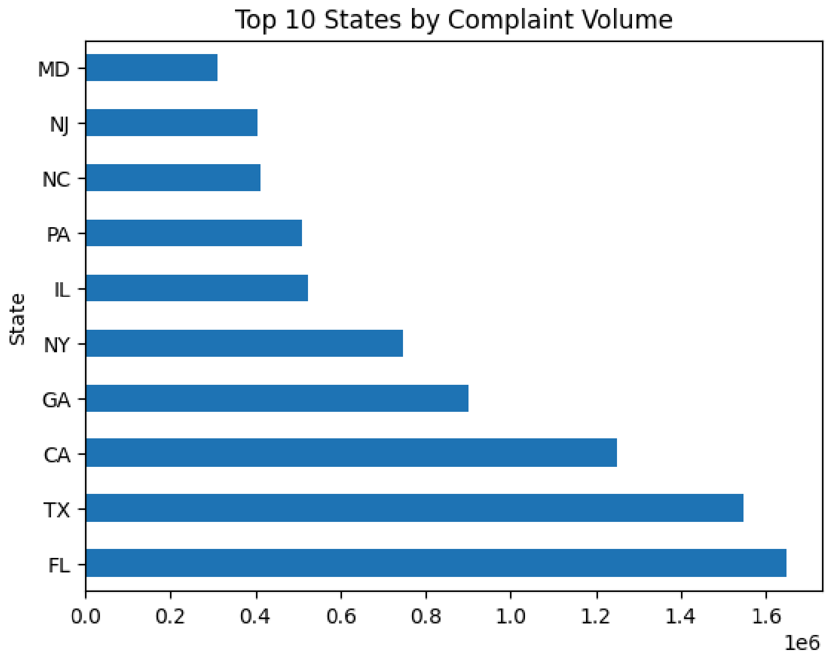




6. Geographic Analysis

Top complaints states:

- CA, TX, FL, NY, GA
  Bar plot highlights CA as the highest.



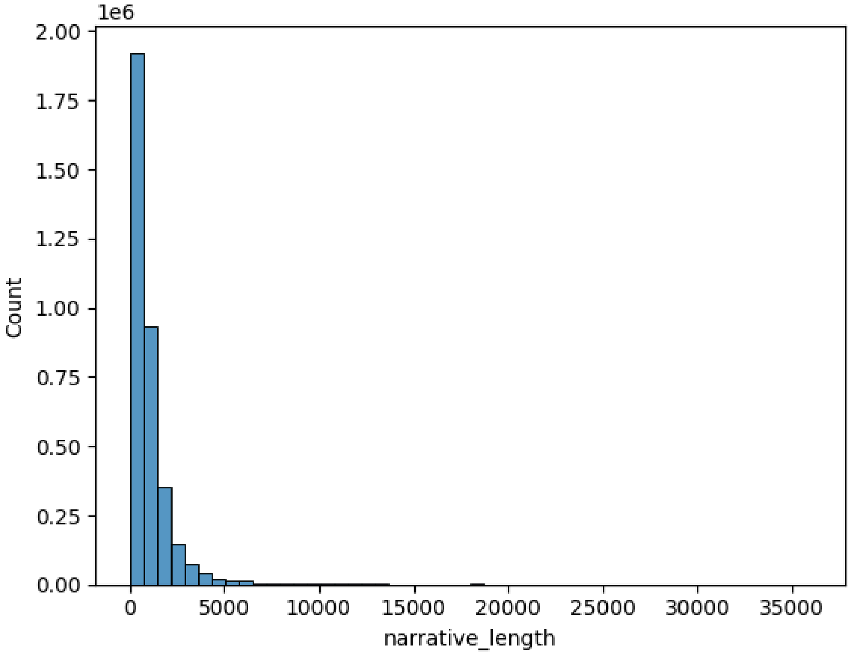
7. Text Narrative Analysis

After dropping missing narratives, the dataset reduces from 12.2M to 3.53M complaints. This filtering is crucial for NLP tasks.

7.1 Narrative Length Distribution

- Mean length: **176 words**
- 75th percentile: **211 words**
- 95th percentile: **515 words**
- Max: **6,469 words**

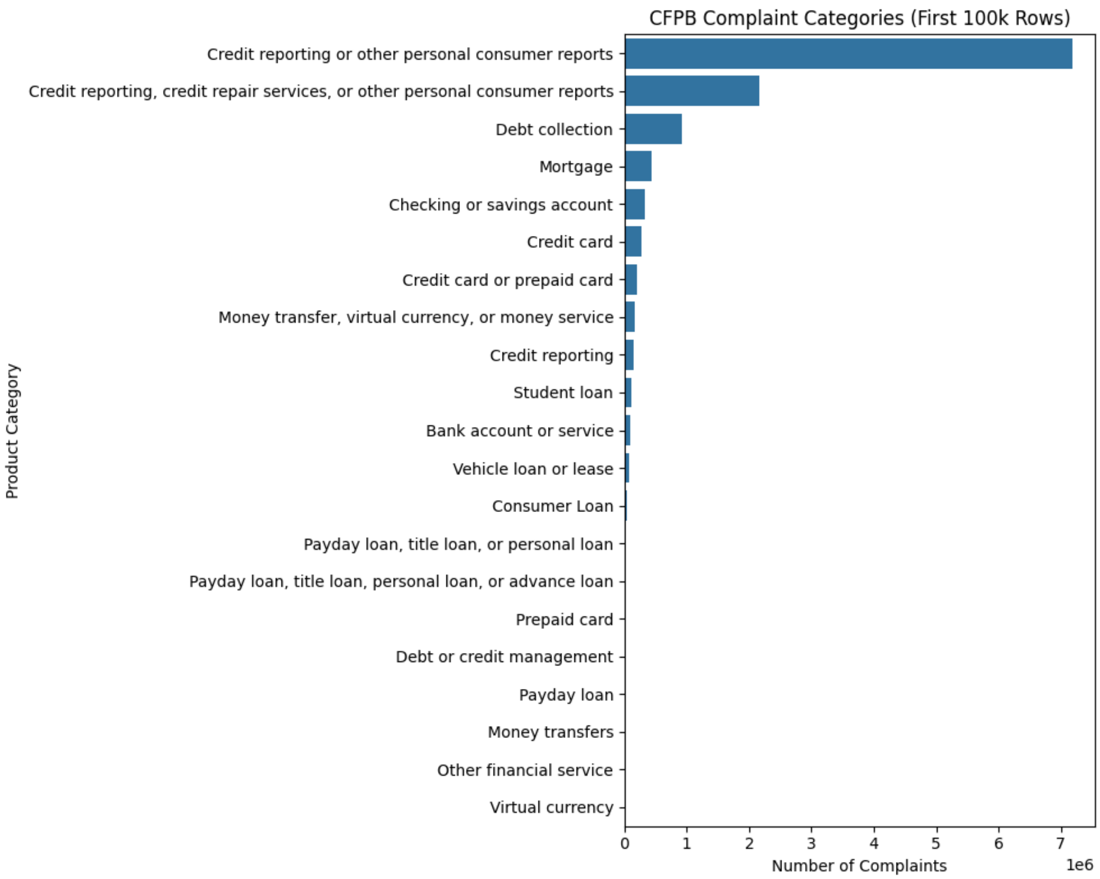
A histogram (with DistilBERT’s 512-token limit) reveals many narratives exceed typical transformer input size, this motivates truncation or hierarchical processing.



7.2 Word Count by Product Category

Boxplots show:

- Credit reporting categories have dense, long narratives
- Debt collection also shows high variance



8. N-gram Analysis

Four categories were compared:

- **Credit reporting variants**
- **Debt collection**

8.1 Key findings

- **All credit reporting categories have identical top n-grams:** xxxx, xxxx xxxx, credit, report, information, account.
- **Debt collection shows:** debt, collection, but still overlaps with credit terms

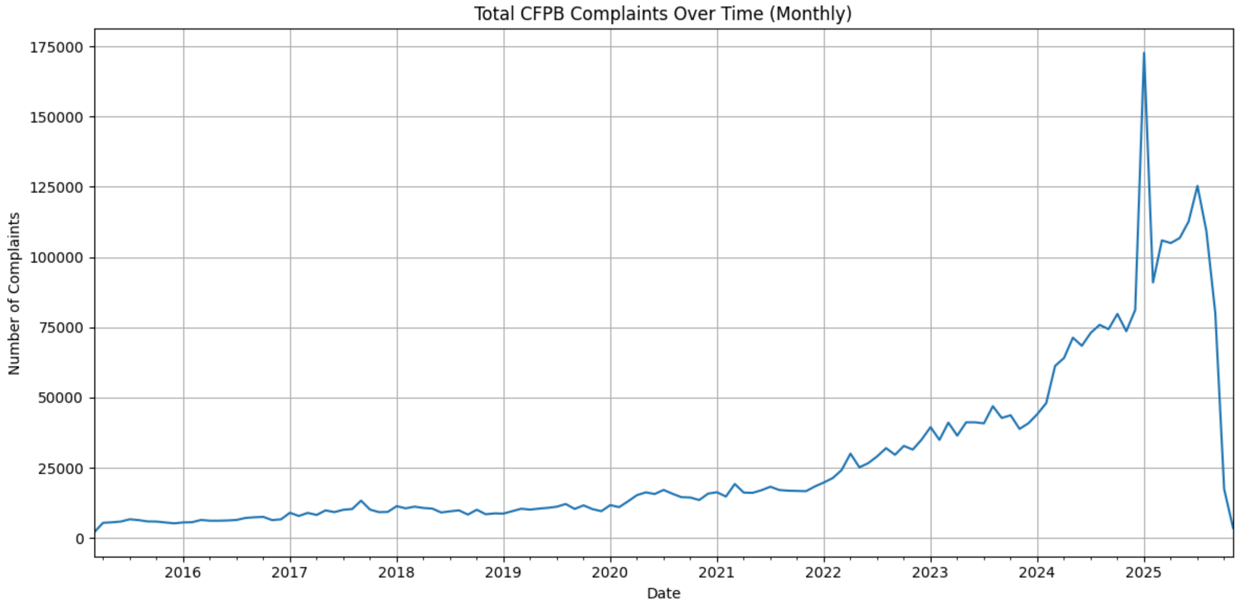
8.2 Insight – Label Redundancy

The notebook concludes that three separate "credit reporting…" product labels are effectively duplicates, evidenced by identical linguistic patterns.

This supports merging them for modeling.

“xxxx” Tokens

These appear due to CFPB redacting personal information (names, numbers) in narratives. This affects preprocessing decisions and model vocabulary.



9. Conclusions

- The dataset is extremely large and text narratives are heavily missing.
- Complaint categories are highly imbalanced.
- Many product categories appear redundant and should be merged.
- Long complaint narratives may exceed transformer limits.

10. Summary

This EDA provides a comprehensive understanding of structure, distribution, quality, and linguistic characteristics of the CFPB complaints dataset. The insights directly support downstream NLP tasks such as product classification, issue detection, and transformer-based modeling.