1.0-final-customer-complaint-eda

December 10, 2022

0.1 Inspection of the pre-processed Data

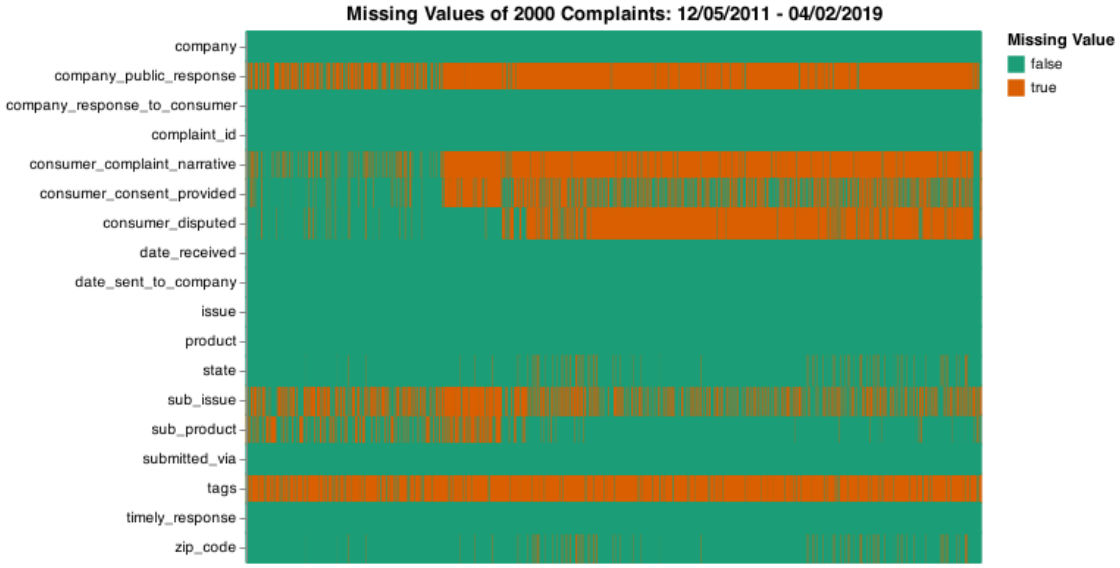
- We can see that the interested target only has 768443 valid values, under which we want to trim the data frame to have null dispute responses removed.
- We can drop non-useful and unique features like `zip_code` and `complaint_id`.
- It seems that we can process the `consumer_complaint_narrative` using NLP and other useful features using `OneHotEncoder` (apply binary encoding if necessary) since the unique values of most of the features are not too many.
- We would also be scaling the numerical features with the `StandardScaler` so that they all have the same range of values centered at the mean.

	columns	valid_count	unique_count
0	date_received	3122836	4018
1	product	3122836	18
2	sub_product	2887543	76
3	issue	3122836	165
4	sub_issue	2438461	221
5	consumer_complaint_narrative	1121913	979279
6	company_public_response	1359325	11
7	company	3122836	6579
8	state	3082743	63
9	zip_code	3082222	34463
10	tags	353109	3
11	consumer_consent_provided	2297533	4
12	submitted_via	3122836	7
13	date_sent_to_company	3122836	3967
14	company_response_to_consumer	3122832	8
15	timely_response	3122836	2
16	consumer_disputed	768440	2
17	complaint_id	3122836	3122836

0.2 Missing Values

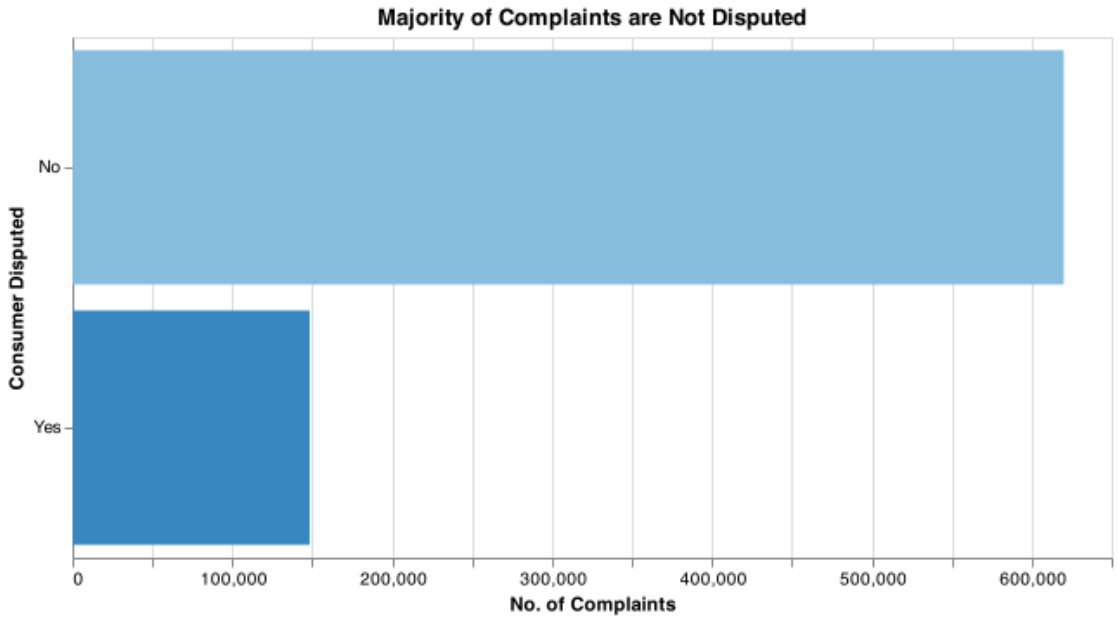
- We also plot the missing values for the top 2000 complaints as a rectangle plot to visualize the distribution of null values for every column in our data.

- We can visually observe the proportion of missing values in our data for the selected number of complaints.
- The orange values indicate the missing values in the data.



0.3 Distribution of Consumer Complaints

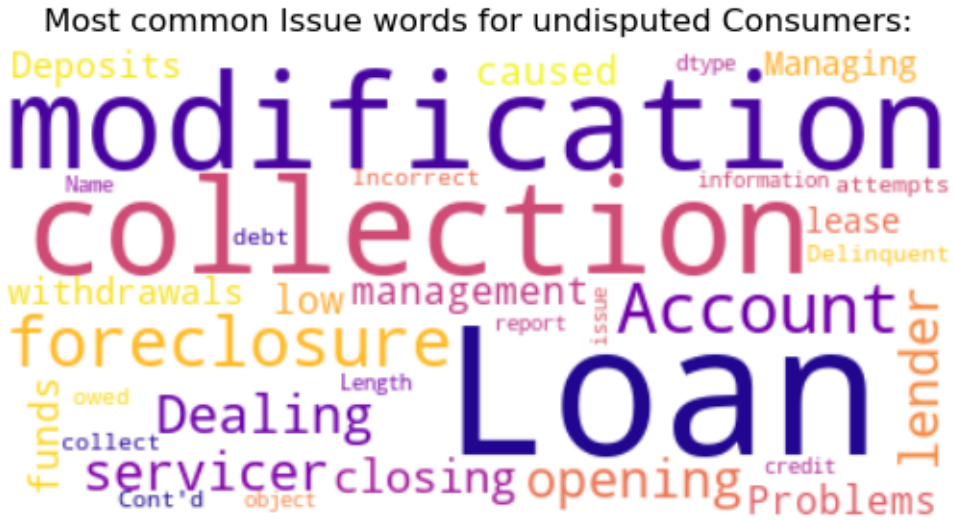
- We also see the number of disputed and undisputed complaints distributed in the dataset
- From the graph below, we observe an imbalanced class, which we should take into account during later training of the model by maybe incorporating different weights and hyperparameters to our model.
- We have a majority of customers who are not disputed comapred to disputed customers in our data.



0.4 Wordcloud of Customer Review

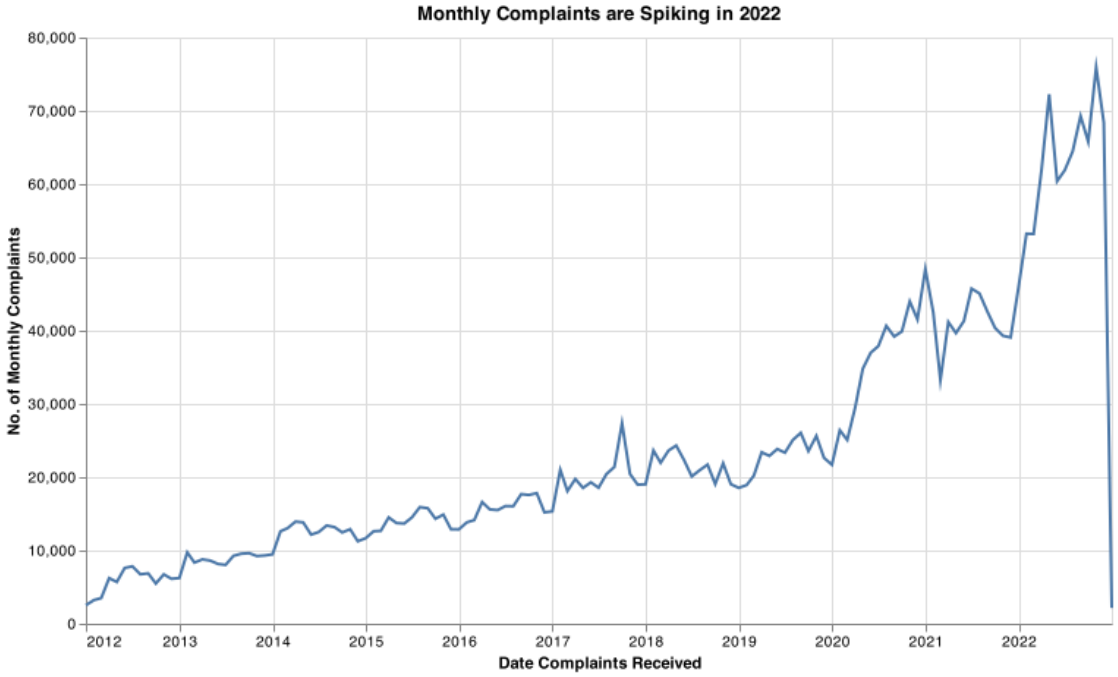
Here we show two visualizations of what customers mentioned for their complaints in disputed/non-disputed classes. We wanted to see the most common words that the customers have mentioned in their narratives when they are disputed with the company's response comapred to the words that they would use when they are not.

`Comparing the Most common words in issues between disputed and undisputed consumers:`

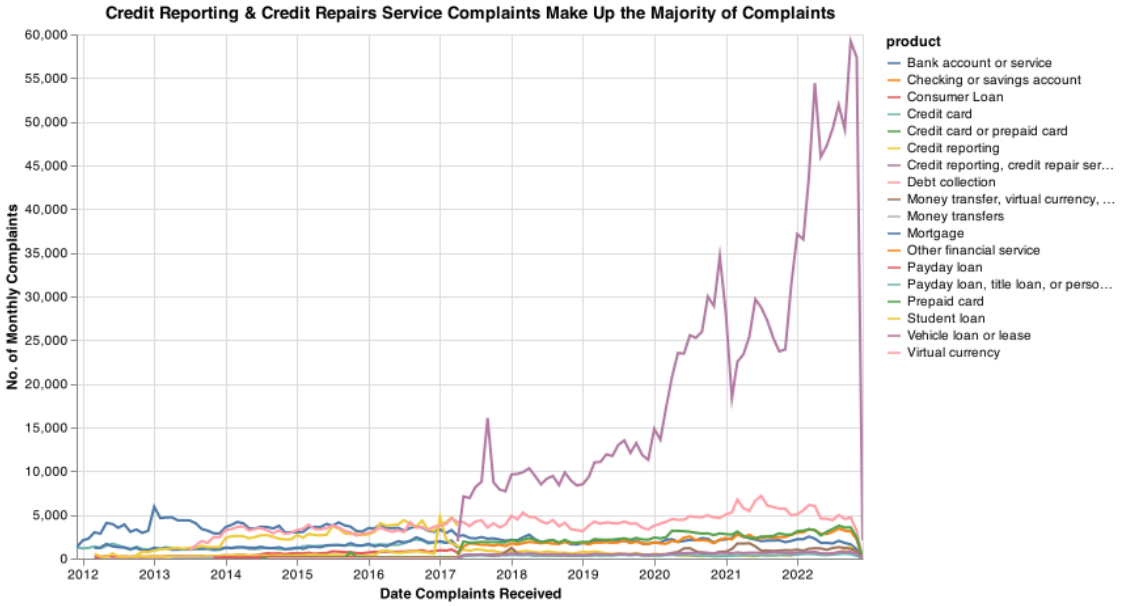


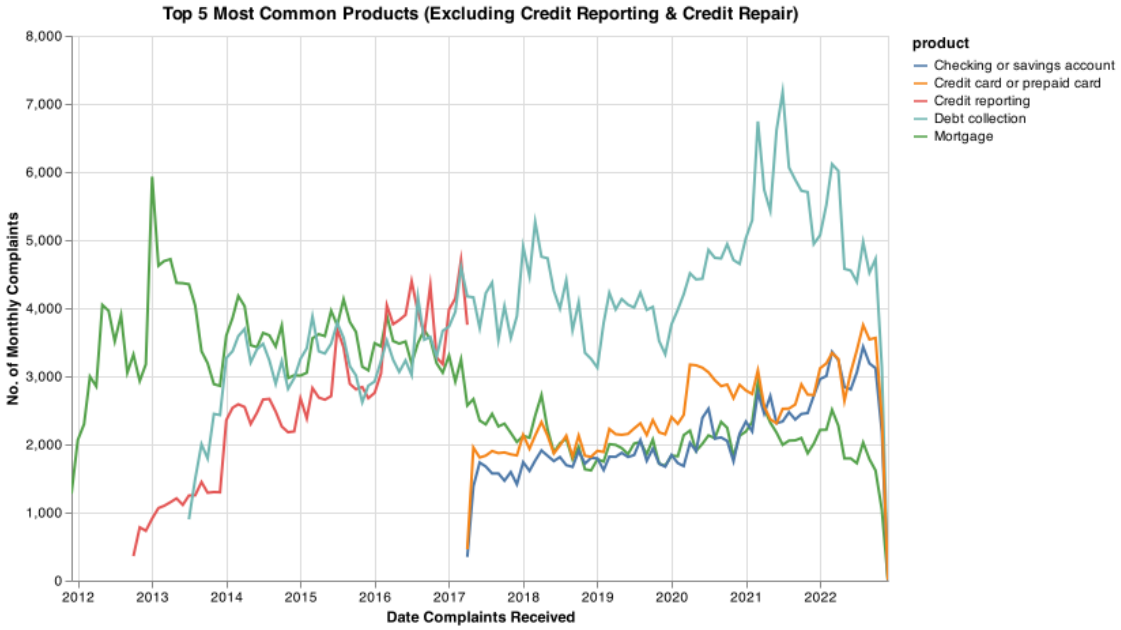
0.5 Insights

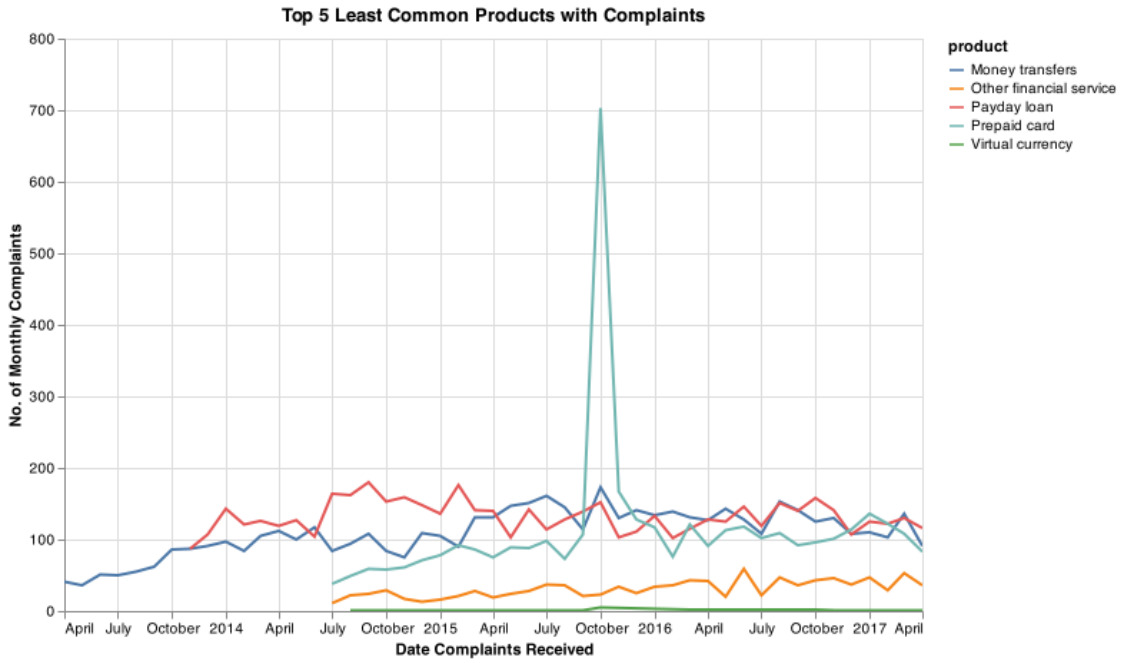
- from July 2022 to November no consumers were recorded as disputing a claim, potentially because they haven't been processed yet?
- what about older claims?



0.6 Complaints by Product



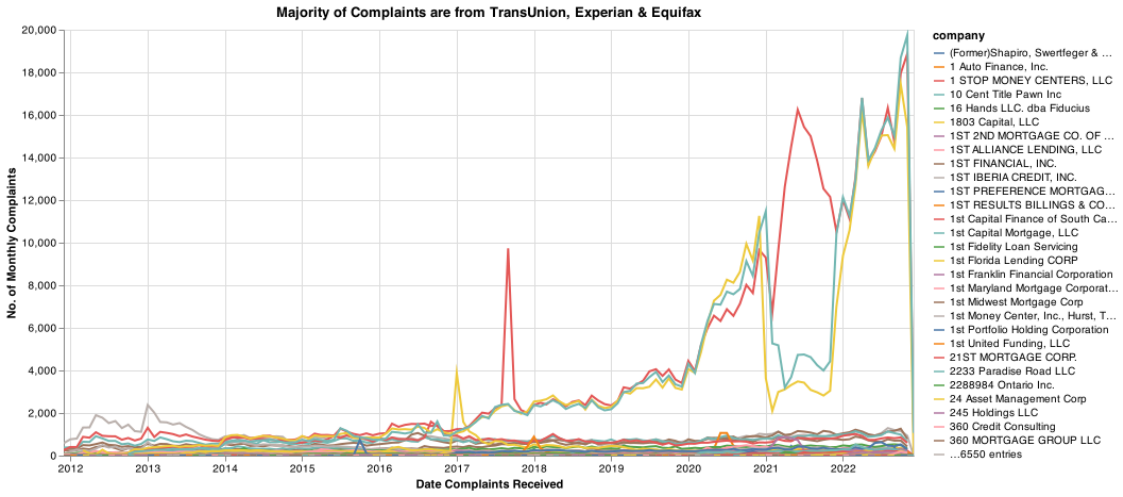




---

1 Complaints by Company

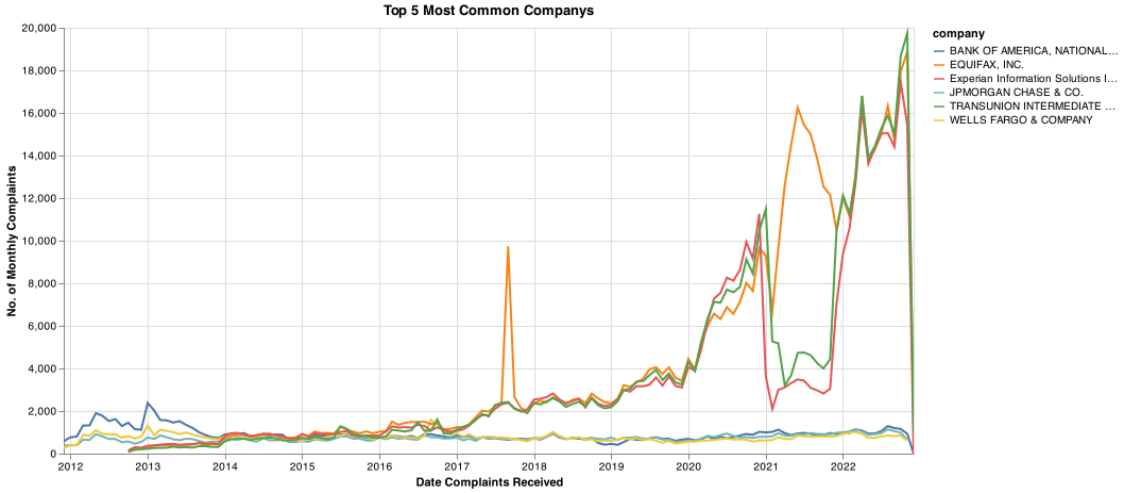
`WARN Symbol legend count exceeds limit, filtering items.`



Insight: looks like majority of compalints come from 3 companies

---

1.1 Top 6 Companies by Number of Complaints



Insight: Top 3 Companies with Most Complaints: Equifax, TransUnion, Experian

---