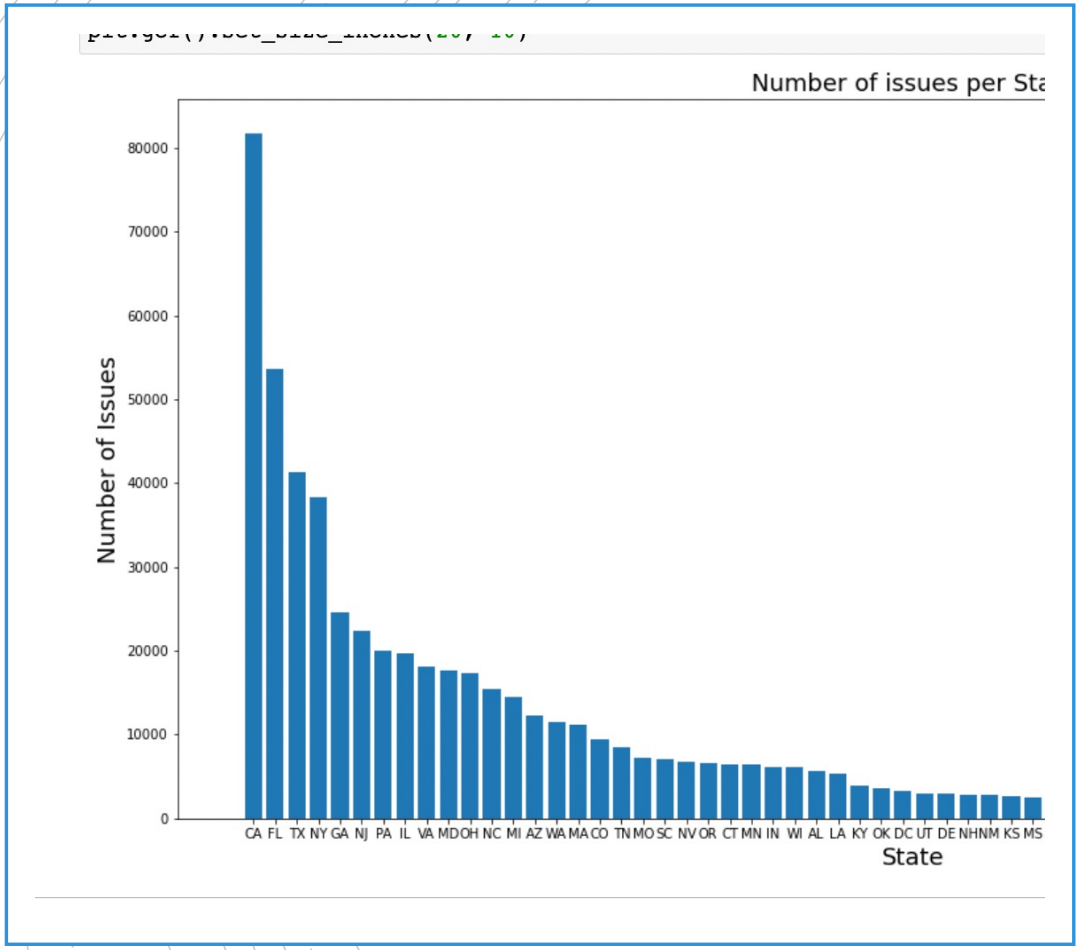
Consumer Complaints Data Analysis

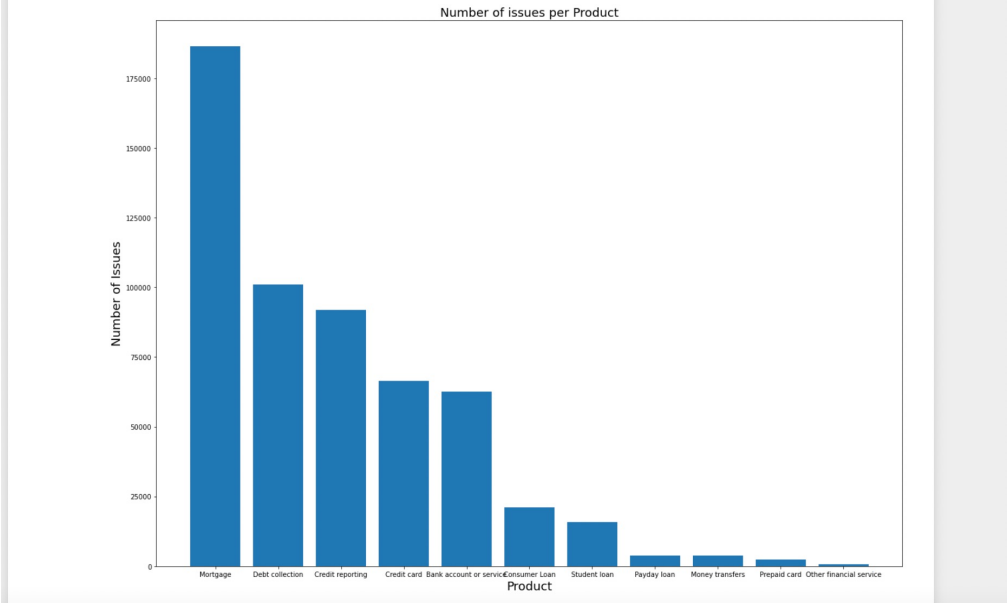
Aditya Kakarla



How does location affect the number of issues?

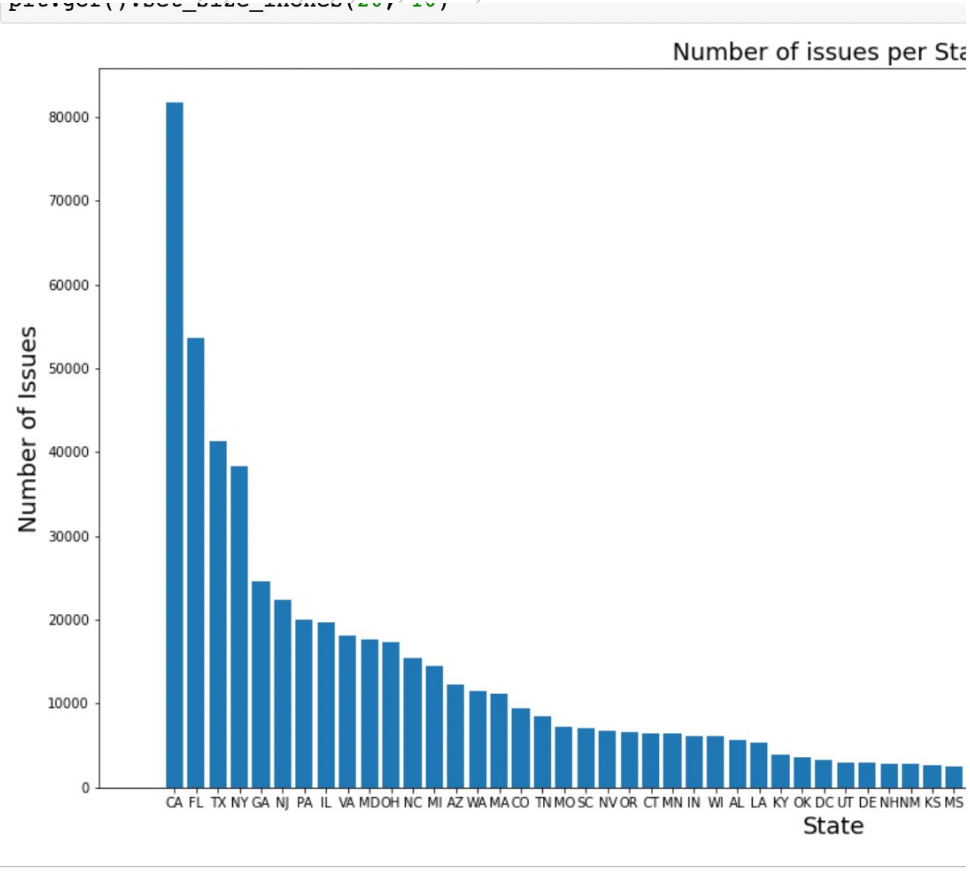
- Based on the data it is clear that California had the most number of disputes
- The way location correlated to the number of issues was that the more populated states(Cali, Texas, Florida) tended to have more issues.
- Moreover it seems the West Coast tends to contain a larger number of issues.
- The insights yield for further analysis could be?

- What regulations'/laws are unique to certain states on the West Coast that cause them to have higher number of disputes?
- What certain demographic subsets are more likely to have disputes?



How does product affect the number of issues?

- From the visualization it is seen the mortgages is the largest source of disputes.
- There was no real subset of products that tended to have a larger number of disputes.
- The insights yield for further analysis could be?
  - What makes mortgages disputed so often?
  - Why is there not real pattern in the subsets of products and the frequency of disputes



How does location affect the number of issues?

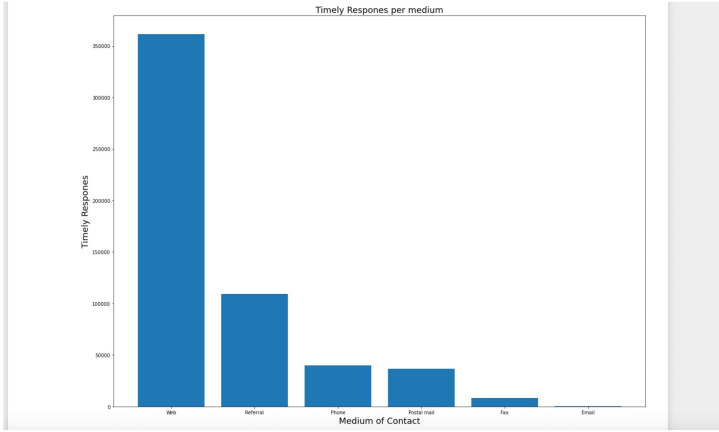
- Based on the data it is clear that California had the most number of disputes
- The way location correlated to the number of issues was that the more populated states(Cali, Texas, Florida) tended to have more issues.
- Moreover it seems the West Coast tends to contain a larger number of issues.
- The insights yield for further analysis could be?

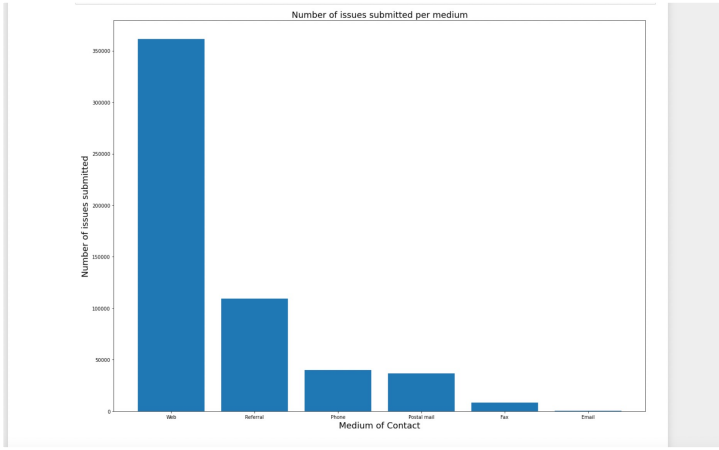
- What regulations'/laws are unique to certain states on the West Coast that cause them to have higher number of disputes?
- What certain demographic subsets are more likely to have disputes?

How does the medium of contact affect the number of issues and timely response rate?

- Online medium of contacts tended to have both a larger number of disputes and a more timely response rate in comaparions tp non-online mediums of contact.
- Direct correlation and match between mediums of contacts that had larger number of disputes and a better timely response rate.
- Moreover, it seems the West Coast tends to contain a larger number of issues.
- The insights yield for further analysis could be?

- What makes digital mediums of contact efficient and inefficient simultaneously?
- How can non digital forms of contact be optimized?





Conclusion

- Further research and exploration regarding the finding correlations between different variable columns in the data set as well as applying more machine learning models would enable us to further visualize the data and finding more insights into the dataset.
- Regarding the Ml models The decision tree model has a lower accuracy but because it deals with the categorical data better it manages to correctly predict some of the disputed issues.
- Since both models are very quick and dirty, detailed parsing of the data using the nltk library would be the ideal direction that should be taken in getting a higher recall and higher number of true positive recognition in the future.
- However, it must be noted that the visualizations and insights provided by the dataset do inform of key areas for our company to be wary and cautious of, and with constant improvement of data preparation and more robust training models or models that offer different sets of insights there lies a higher scope of predicting disputed issues and even company responses.