Benchmarks:

An Analysis of Massachusetts Court Reversals

CS 506 Final Project Report

Professor Adam Smith

Spark! Advisor: Professor Maggie Mulvihill

Spring 2018

Authors:

James Craver

Sam Harrison

Haley Pereira

Code and data are available on github

1. Introduction

We are working with *BENCHMARKS: A Citizen's Scorecard on Judicial Accountability in Massachusetts* to determine the conditions under which judicial rulings are more or less likely to be reversed. Ideally, the project will uncover never before-seen patterns of judicial behavior and decision-making in criminal and civil cases, which will have the effect of recasting and spurring more aggressive local news coverage of the state court system. As a result, the project should become a valuable tool for judicial accountability. This project is mentored by Maggie Mulvihill from the College of Communication Journalism Department and is partnered with the Boston Globe. Benchmarks seeks to combine data analysis and investigative journalism to create results that are more impactful than just using one of the two. Specifically, this project should serve as a building block for a new Boston University interdisciplinary news team which will examine case reversals in all 50 states over the past decade.

Our work on *BENCHMARKS* this semester focuses on criminal cases appealed to the Massachusetts Supreme Judicial Court and Appellate Courts from 2008–2018. We want to identify similarities between cases that are reversed, as well as identify what factors determine whether or not a case is reversed. This analysis will serve as a starting point for further analysis, such as analyzing civil cases, or going back further into the past.

2. Questions

- What proportion of cases are reversed in Massachusetts?
- What are the similarities between cases that are reversed?
  - Are there certain types of crimes that are reversed more than often than others?
  - What factors are most strongly correlated with the reversal of a criminal conviction?

3. Data

Our first set of data was annual reports of court statistics. We used the Massachusetts Judicial System Annual Reports (see the 2017 report here) as part of our initial data analysis and to get an understanding of how the court system in Massachusetts works. Only reports from 2009–2017 were available. The 2018 Report has not been released because FY2018 is not over yet, and reports are not available from before 2009. These reports are publically available and did not require any special techniques to gather; we just downloaded the PDF files and read them.

Our second dataset was case summaries from the appellate court system. We scraped the Massachusetts Appellate Court Website here in order to make a determination about whether each case was reversed or not and to gather identifying information about each case. Each page contains the parties involved, the nature, title, and status of the case, the counsel present, the lower court who initially decided the case, and the list of docket entries detailing the lifecycle of the case. We considered this information to be foundational in the attempt to document the state of the appeals court, which is itself a reflection of the lower court system (which is not subject to any standard of documentation).

The challenge we faced in gathering the data was threefold: how to find a full list of cases, how to download all of them for processing, and what to do with them once we had downloaded them. To overcome the first challenge, rather than relying on the limited searching functionality available in the site and attempting to parse links from those results, we identified the pattern in page names and exploited the sequential numbering to generate a list of URLs to consider for web scraping. There are 4 types of cases we considered: J (justice), P (panel), SJ (single justice), and SJC (supreme judicial court), and since each is numbered according to a predictable scheme, we were able to identify the largest docket number for each year we considered and generate the URL with limited interaction with the website. This reduced the overall number of queries required.

Next, we had to download all the cases found at the URLs we had generated. To accomplish this, we used the Python Requests module and requested each case one at a time, randomizing the number of seconds between each request and the user agent we added in our request header. We also limited the number of downloads per IP address, utilizing VPNs to increase our downloads per day, and limiting the number of downloads at a time in order to not over strain the MA Appellate Court website. We also had the download module check if a page were already present already before making the request and skipping that request if we already had downloaded that page. This way, in the case of a failed download, the user simply needed to regenerate the URLs and delete the incorrect files and rerun the download without fear of duplicating work. Using this, over the course of weeks we were able to download all cases from the MA Appellate Court system between 2008 and 2018. This code is available on our github repository in the Jupyter Notebook with the name of `CS_506_-_Spring_2018_-_Benchmarks_Data_Download_Module.ipynb`, and the resultant HTML files are available as a zip archive on Google Drive here.

Finally, once we downloaded all the case summaries, we began our main analysis notebook, entitled `CS_506_-_Spring_2018_-_Benchmarks_Data.ipynb`. To run this code, unzip the archive linked to in the previous paragraph into the same directory as the git repository and then do a “Run All” on the Jupyter Notebook (This will take a while to run; it took us about 30 minutes but will vary). This notebook contains the code to read all the downloaded HTML and convert it into a lists of dictionaries for processing (these lists were also printed out as the CSV files cases.csv and dockets.csv, which are available in our github repository). To accomplish this, we used the HTML parsing package Beautiful Soup to parse the name of each bolded tag on the page and the value next to it, interpreting these as column titles and entries. The docket entries were parsed separately, with each docket entry being recorded as the case ID, entry date, and entry text. Thus each case could be converted to a dictionary of key value pairs, including case id, title, nature (criminal or civil), status (e.g. closed), plaintiff, defendant, relevant dates, and a list of docket entries. We also added two new columns: Has Affirm and Has Reverse which indicate that the docket entries contain the string "affirm" or "revers", respectively (the latter to capture both "reversed" and "reversal"). This allowed each case in the appellate court system to be represented as a dictionary object which could be used for later analysis. For the scope of this project, we limited the analysis to cases whose type was Criminal and status was Closed: Rescript or Decided: Rescript (indicating the appellate court had decided the cased and issued its decision back to the lower court).

Our third source of data was the full opinion text. In order to get the full opinion text, we downloaded data from the Massachusetts Appellate Opinion Portal here. This website has fairly sophisticated anti-scraping techniques applied to it, so we manually extracted the data. Because of this, we decided to only download the data from reversed cases, as doing this process manually for every appealed case would take too long. An additional concern with this dataset is that we are only able to download cases from the Supreme Judicial Court with published opinions, which only accounts for about 10% of all reversed cases. The other 90% of cases do not have published opinions and so it was not possible to download any opinion text for them.

4. Analysis

4.1 Exploratory Analysis

The data in Table 1 was compiled from the Massachusetts Judicial System Annual Reports. We then used this data to create a bar graph showing the rate visually by year (see Figure 1). This table helps us visualize our data and determine the reversal rate of appealed cases. Table entries where data is not available were left blank.

Table 1: Reversal Rate Data

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total decisions		771	765	889	750	678	686	709	728	734	
Decision of lower court affirmed		636	524	653	592	550	572	603	581	594	
Decision of lower court reversed		91	196	150	109	89	77	67	99	102	
Other result reached		43	45	86	49	39	37	39	48	38	
Reversal ratio		0.118	0.256	0.169	0.145	0.131	0.112	0.094	0.136	0.139	
Published opinions						69	64	71	87	83	
Summary dispositions						609	622	638	641	651	

Publishing ratio						0.102	0.093	0.100	0.120	0.113	

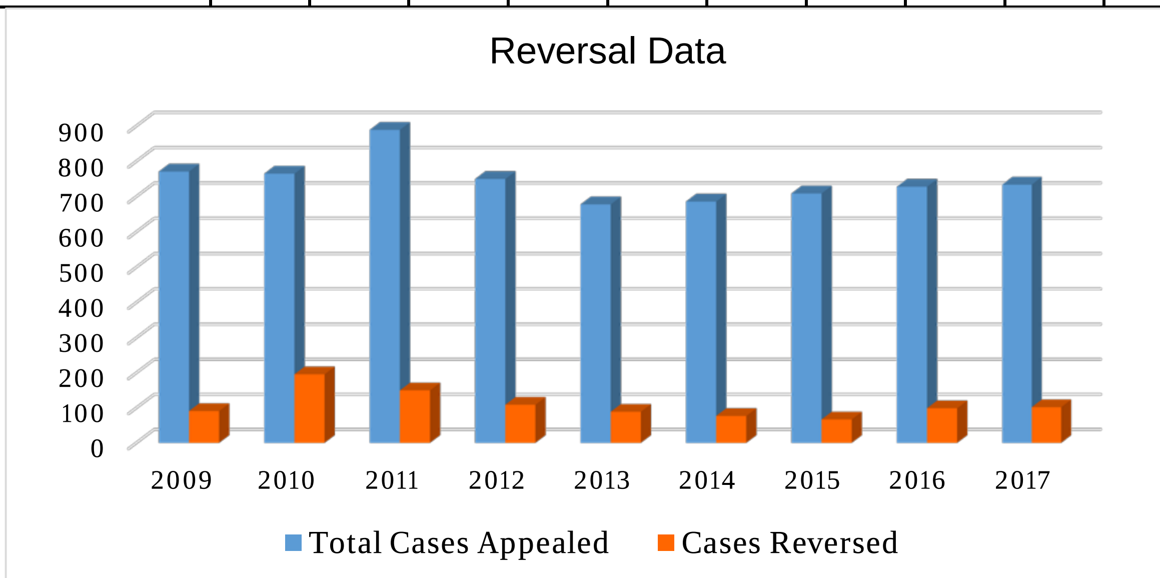

Figure 1: Graphical View of Reversal Rate

4.2 Analysis of Case Summaries

We attempted two separate analyses of the case summaries: analyzing the case metadata and analyzing the docket entries.

Within the metadata, each case lists the lower court judge who decided the original case, and since a reversal would indicate a correction to the lower court judge, we were able to find total cases appealed and total cases reversed by judge and find judges with the highest reversal percentage. This would not take into account cases which a judge decided that did not make it to appeals, which may not be fully available in our data set, so we limited this analysis to closed cases.

We found that the median percentage of reversal percentage by lower court judge was 9.5%, but the mean reversal percentage was 15%, so we concluded that there are a small set of judges who tend to have a higher percentage of reversals. With that notion, we identified a number of judges who had a much higher percentage. Excluding judges with fewer than 5 cases (which includes judges who have had all of their small number of cases reversed), we identified the below 10 judges who have a reversal percentage of 50% or greater.

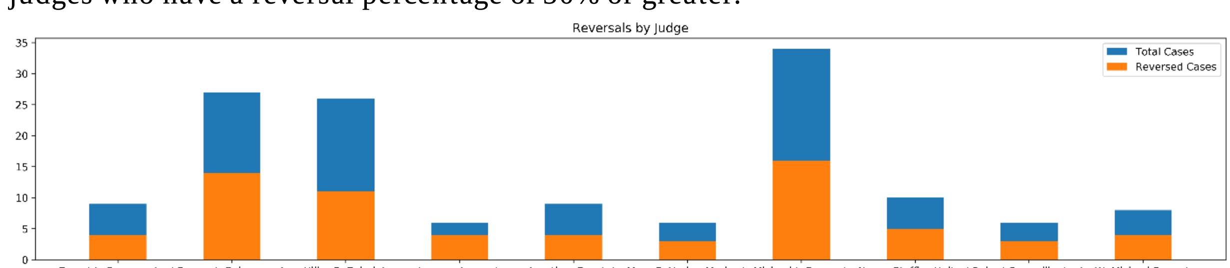

Figure 2: Lower Court Judges with Highest Reversal Percentages

Another avenue we pursued was to see whether the number and length of extensions requested by each side correlated with reversal. Training a decision tree by total time of the appeals process, number and length (days) of extensions requested by the prosecution, and number and length (days) of extensions requested by the defense (training on 80% of the data), we found that the area under ROC curve of the predictions on test data was 0.47, so we conclude that, despite our intuition, time taken by either side is not sufficient to predict outcome.

We also looked at the docket entries. At a high level, what we wished to do was to determine how similar different docket entries were to provide some insight into how each appeal proceeded and whether any type correlated with reversal. We chose cosine similarity as our measure for clustering, as this technique is frequently used in measuring similarity of term frequency.

To prepare the docket entries for analysis, we used a tf-idf measure in order find the most meaningful terms in the case, using sci-kit learn's TFIDF Vectorizer. Care was taken not to cluster by time, so we removed all numbers from the data before applying tf-idf score. We used the monograms, bigrams, and trigrams found in each docket entries because legal terms are often more than one word; this is available as a parameter in the aforementioned module. Using Sci-Kit Learn's TruncatedSVD module, we transformed all docket entries, revealing 4 principal components; beyond that, the subsequent terms offered much less meaning per dimension. Now that this data was available in numerical form, we were able to use SciPy's hierarchical clustering module to link all transformed docket entries using the cosine similarity. By measuring the silhouette score of clustering into between 2 and 15 clusters, we identified 4 clusters as the most meaningful number of clusters to use. The graphical representation we found supports this conclusion.

The figures below demonstrate the results of transformation and clustering. The first diagram shows each cluster that was found by the process described above, labeled by most common terms in each cluster. The final label is the sample representative of each cluster and represents its location on the graph.

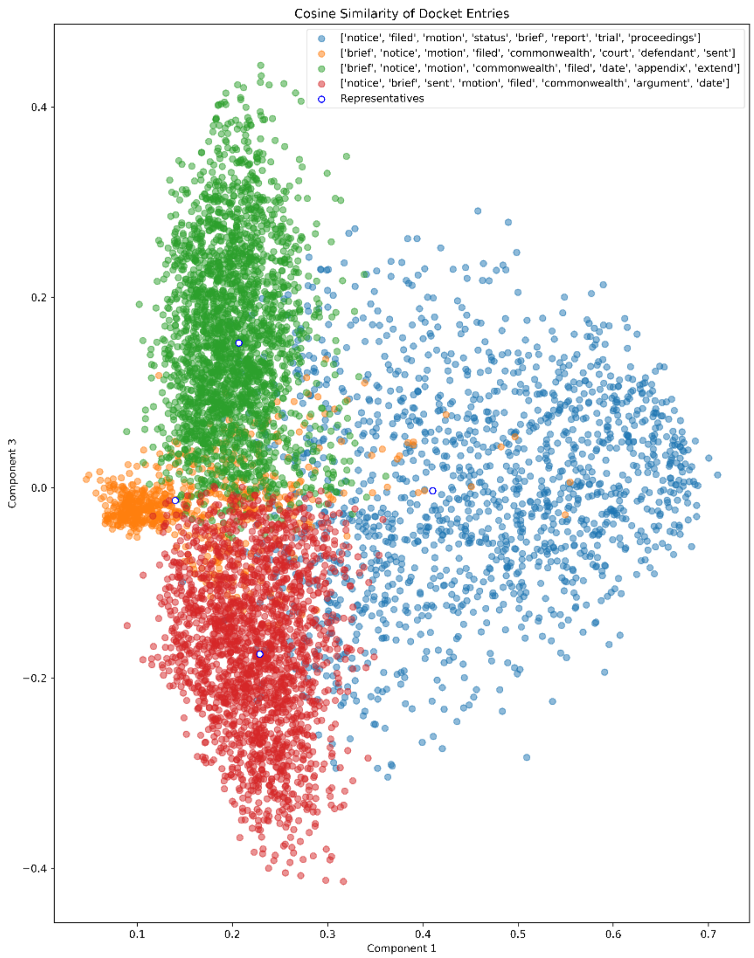

Figure 3: Clustering of Docket Texts

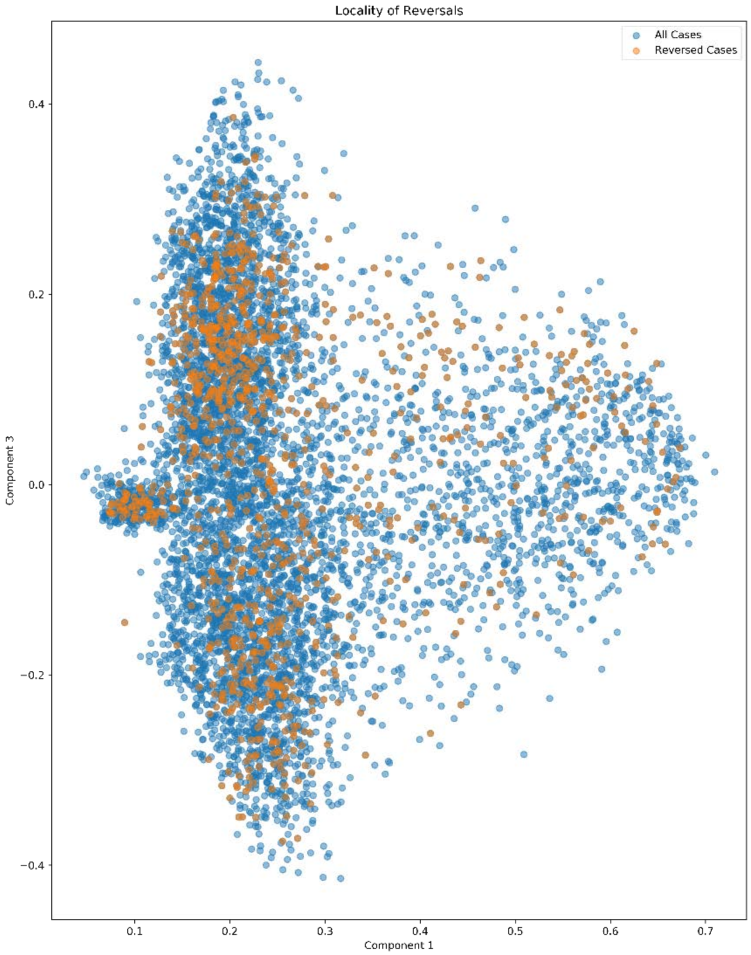

Figure 4: Clustering Colored by Reversal Status

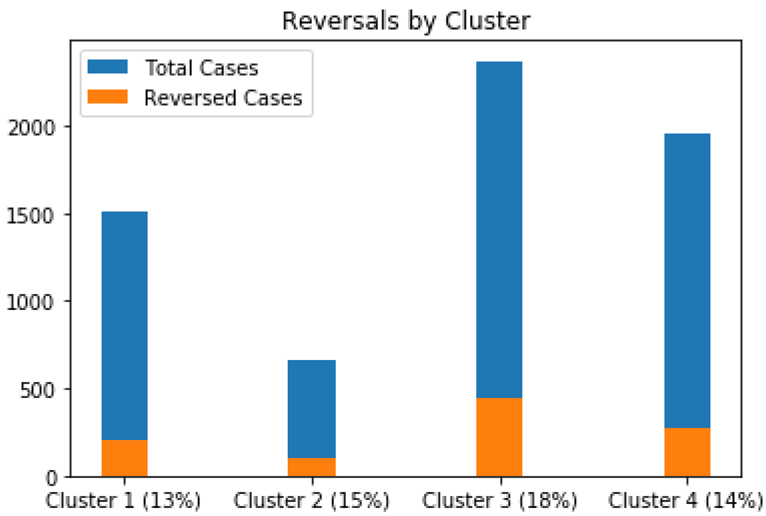

Figure 5: Reversal Percentage by Cluster

Determining the meaning of each cluster was a challenge. At the surface level, relative term frequency between the clusters indicates that cluster 1 has more notices and motions than briefs, clusters 2 and 3 have more briefs than notices and motions, and cluster 4 has more notices and briefs than motions. To elucidate the meaning behind each cluster, we found the sample representative of each cluster. To do this, each case within a cluster was given a score of intra-cluster similarity by summing its cosine distance to all other cases within the cluster, with a higher score indicating that it is similar to more cases (since the cosine similarity of tf-idf scores ranges between 0 to 1, the range of this score would be between 0 and the number of cases in the cluster). Even though cosine distance does not satisfy the triangle inequality, finding the maximum score in each cluster found a case that was centrally located, and these representatives have been added to the clustering graph as a fifth label. Using this representative and other high scoring cases, we were able to determine the following profiles of each cluster:

- Cluster 1: This cluster contained the third highest volume (~1500 cases) and had the lowest reversal rate at 13%. From the docket entries of the high scoring cases, we determined that proceedings in these cases were stayed (as in the court halted proceedings), whereas other clusters' cases did not feature this.
- Cluster 2: All Supreme Judicial Court cases ended up here; docket entries typically contained more detail than other cases. This cluster, consistent with expectation, contained the lowest volume (around 500 cases) and had a 15% reversal rate.
- Cluster 3: This cluster contained the largest volume of cases (~2500 cases) and also had the highest percentage of reversals at 18%. Even though cases were stripped of numerical data beforehand, cases in this cluster tended to be from the years 2008-2011, indicating a standard of wording that was used more often during this time. We attribute the larger number of reversals to the fact that 2010 contained an abnormally large proportion of reversals compared to other years.
- Cluster 4: This cluster contained the second highest volume (~2000 cases) and had a 14% reversal rate. Cases in this cluster tended to be from the years 2012 onward.

4.2 Analysis of Opinion Texts

We analyzed all of the Opinion Texts that we were able to download in order to determine the proportion of reversed cases by the type of crime. In order to do this, we searched the headnotes of each opinion page for keywords such as “homicide” and “firearm.” By looking for these keywords we were able to classify the reversals based on the type of crime. Figure 6 shows the results of this analysis in a graphical format. This analysis provides a starting point in determining what factors in a case make it more or less likely to be reversed. However the exact search criteria could be improved to get a more sophisticated view of the reasons for reversal.

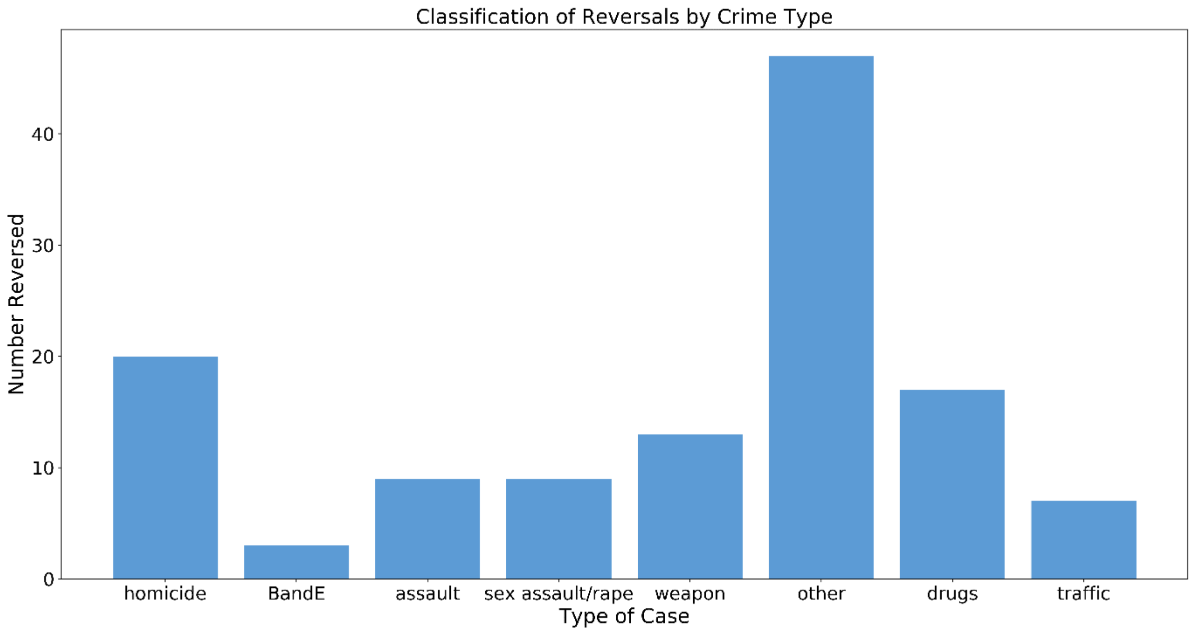

Figure 6*: Reversals by Crime*

5. Conclusions

We were able to find the reversal rate of criminal cases in Massachusetts. This is about 14%, although the rate varies from year to year (see Figure 1). Most years were between 10% and 15%. 2010 was noticeably higher with a reversal rate of 25.5% and may be an outlier skewing the trend.

Analyzing the case summaries found on the MA Appellate Court website, we measured a number of aspects of each case, including lower court judge, extensions filed by either side, and textual analysis of docket entries. Intuitively, we expected time taken by either side to correlate with reversal, but this did not appear sufficient based on our analysis to predict a reversal. Likewise, parsing the docket entries proved mainly as a tool to classify the proceedings, and though it indicated a small increase in reversals correlating with whether the court allowed a stay of proceedings, this increase is not significant. The largest indicator we found was based on lower court judge; we identified a number of judges with a high percentage of reversal, ranging around 50% for judges with a significant number of cases, so we conclude that lower court judge is a significant predictor of reversal and that such ought to be a starting point for further investigation as to why that would be the case.

From our analysis of the opinion texts, we see that homicide is the most likely crime to be reversed by the Appeals Court, followed by drug and weapon-related crimes. This conclusion needs to be qualified, however, because we are basing this only on reversed cases with published opinions. We may not have enough information to fully answer the question of which factors correlate with reversal, or which crimes are reversed most frequently. Nevertheless, it appears that the crime type has an effect.

6. Future Steps

To do a more thorough analysis of this data, we would like to incorporate more potential factors in whether or not a case was reversed, like data about the relevant police department, or demographic information about the judge, defendant, and attorneys. To do this, we will need to find additional data sources, but this initial analysis as well as the data recorded from the MA Appellate Court website should give us (or future teams) a foundation on which to generate further analyses. For example, one could perform a regression on the case metadata as a whole. The metadata about the case may not be useful in predicting the outcome, but it could generate some key predictors which would indicate a bias in the judicial system that would merit further investigation. Another way the metadata could be useful is as a springboard to find other data sources. It will provide a way to lookup other key case characteristics on other data sources.

Another improvement we could make to our analysis would be to obtain a more expansive source of opinion texts. With our current data source, we are limited to cases with published opinions. It would make for a much more robust analysis if we could get access to the unpublished opinions. This, however, would require petitioning the court system and would likely be subject to large amounts of government bureaucracy. Professor Mulvihill is working on finding a way to get access to a broader range of this data, but this is unfortunately not likely to happen within our project timeline.