CS 506 FINAL PROJECT REPORT

BENCHMARKS:- AN ANALYSIS OF MASSACHUSETTS COURT REVERSALS

Instructor - Professor Lance Galletti

BU Spark! Advisor - Professor Maggie Mulvihill

Technical Advisor - John Merfeld

Fall 2019

AUTHORS:

Abhimanyu Jain

Anton Njavro

Yashvardhan Das

Yuwan Xiao

Code & Data are available on GitHub

1. INTRODUCTION

This project is based on a continuation of the work done by the previous team during the last year. The work associated is in collaboration with BU Spark! and mentored by Ms. Maggie Mulvihill, Associate Professor of the Practice of Computational Journalism, of the College of Communication at Boston University. A partner associated with this project is The Boston Globe. The primary motive of implementing this project involves analyzing the rate of reversals associated with court judges and determining the factors associated with the types of courts based on these reversals. This project, which is a combination of data retrieval, its subsequent analysis, and investigative journalism, can serve as a foundation for a probable interdisciplinary team at Boston University which can scrutinize case reversals in all the entire 50 states present in the country.

The work done by the previous team last year focused on criminal cases appealed to the Massachusetts Supreme Judicial Court (SJC) and Appellate Courts from 2008 - 2018. Our work on this project in the current semester focuses on civil cases appealed during the same 10-year period. Additionally, a secondary investigative analysis is performed on both civil and criminal cases of the period of 2018 - 2019. We aim to associate probable similarities between the cases that have been reversed and pinpoint certain judges whose reversal rates are considerably high.

2. DATA COLLECTION & CHALLENGES FACED

The data for our analysis were collected through a combination of different websites. The prime challenge we faced was the irregular structure of the pages that needed to be web-scraped. For some unknown reason, the core structure of the web-pages of cases dating from 2018 - 2019 was significantly modified as compared to the main structure of the web pages dating from 2008 - 2018.

As a consequence, we found it extremely difficult to build upon the code-base generated by the previous team. Moreover, this time, the

information was gathered from 2 different websites, namely - http://masscases.com/ and https://www.ma-appellatecourts.org/. The former website was used to generate the decision of the cases and the latter website was used to scrape the name of the respective judge. Many-a-times, the decision version had to be manually understood and then put in the dataset owing to the fact that there is no uniformity in the written structure of the judgment. There is a high level of non-uniformity in the semantic structure of the judgments that have been written. Another noteworthy obstacle we encountered was that some of the names of the judges were not mentioned (provided the docket number was entered) in the pages pertaining to the second website mentioned above. In such cases, we were compelled to enter the name of the corresponding judge as 'Not Mentioned' in the constructed dataset.

Our first major dataset contains the data of civil cases pertaining to the period of 2008 - 2018. This has been taken as a subset of the overall dataset that was used by the former team. To make a meaningful analysis of the data, during the pre-processing stage we had to discard a number of columns present in the dataset. The second major dataset that is used in our project contains details of both the civil and criminal cases relating to the period of 2018 - 2019. Since the court website has sensitive information, there were error requests encountered during the scraping process. Strategic time-limits had to be written in the code to successfully download all the required data.

3. EXPLORATORY DATA ANALYSIS & FINDINGS

The main component of our analysis composes of data-frames generated by the web-scraping of the respective pages. We have utilized this data to generate pie-charts and histograms for visualizing the rate of reversals of judges. A sample subset of one of the data-frames is shown below.

	cases	headnote	text	type	caseid	judge	case status
326	care and protection of m.c.	sjc-12339impoundment. minor, care and protect...	in this case, we consider the appropriate stan...	criminal	sjc-12339	Barbara A. Lenk, J.	Reverse
329	j.h. vs. commonwealth.	sjc-12395juvenile court, delinquent child. pr...	a single justice of the county court reserved ...	criminal	sjc-12395	Barbara A. Lenk, J.	Affirm
332	exxon mobil corporation vs. attorney general.	sjc-12376attorney general. consumer protectio...	in 2015, news reporters released internal docu...	civil	sjc-12376	Heidi E. Brieger, J.	Affirm
336	jane doe no. 1& othersvs. secretary of educat...	sjc-12275education, charter school. education...	five students who attend public schools in the...	civil	sjc-12275	Heidi E. Brieger, J.	Affirm
348	commonwealth vs. marcelo almeida.	sjc-12179homicide. evidence, prior misconduct...	the defendant, marcelo almeida, stabbed the vi...	criminal	sjc-12179	Thomas F. McGuire, Jr., J.	Affirm
352	commonwealth vs. keith cawthron (and three c...	sjc-12322controlled substances. constitutiona...	in this case, we consider whether police offic...	criminal	sjc-12322	Kenneth W. Salinger, J.	Reverse
354	commonwealth vs. jose torres.	sjc-12374stalking. compensation of victims of...	in this appeal, we consider whether a defendan...	criminal	sjc-12374	Heidi E. Brieger, J.	Reverse

Fig. 1: Data-frame constituting information about cases pertaining to the period of 2018 - 2019.

The last column present in the above table depicts the judgment of the respective cases. In this case, there are 3 categories - the decision was affirmed (‘Affirm’), the decision was reversed (‘Reverse’) and, a part of the judgment was reversed (‘Partially’). The next step in our analysis was to sort the names of the judges in descending order of the number of cases associated with them during the respective period of 2008 - 2018. In this case, we have done subsequent analysis on only those judges whose count of cases during the last year and a half is equal to at least 4. This was done keeping in mind the fact that when a very small number of cases are pertaining to a certain judge, the rate of reversal/affirmation can be quite high but at the same time that can be somewhat misleading. For instance, consider a judge “A” has only 1 case associated with him/her. If the case was reversed, then the reversal rate would be 100%. This would bring a very negative connotation for the respective judge who has just administered a single case. Hence, from a statistical point of view, we have selected the minimum threshold regarding the number of cases to be 4. The table shown below shows the sorted order.

	Reversed	Affirmed	Partial	Total
David A. Lowy, J.	0	13	1	14
Frank M. Gaziano, J.	1	8	1	10
Scott L. Kafker, J.	0	9	0	9
Elspeth B. Cypher, J.	1	7	0	8
Janet L. Sanders, J.	0	4	2	6
Richard J. Carey, J.	2	4	0	6
Heidi E. Brieger, J.	2	4	0	6
Douglas H. Wilkins, J.	3	2	0	5
Maynard M. Kirpalani, J.	0	4	0	4
Gary A. Nickerson, J.	0	4	0	4
Kenneth W. Salinger, J.	3	1	0	4
Christine M. Roach, J.	2	2	0	4
Peter M. Lauriat, J.	1	3	0	4
Janet Kenton-Walker, J.	2	2	0	4
Linda E. Giles, J.	0	4	0	4
Daniel M. Wrenn, J.	1	2	1	4
Kimberly S. Budd, J.	0	4	0	4
Thomas F. McGuire, Jr., J.	1	3	0	4
Barbara A. Lenk, J.	2	2	0	4

Fig. 2: Table depicting a sorted list of the number of cases pertaining to specific judges during the period of 2018 - 2019.

For a more in-depth exploratory analysis, we have constructed histograms for each category of decisions/judgments (Affirm, Reverse and Partially). The corresponding visualizations are below.

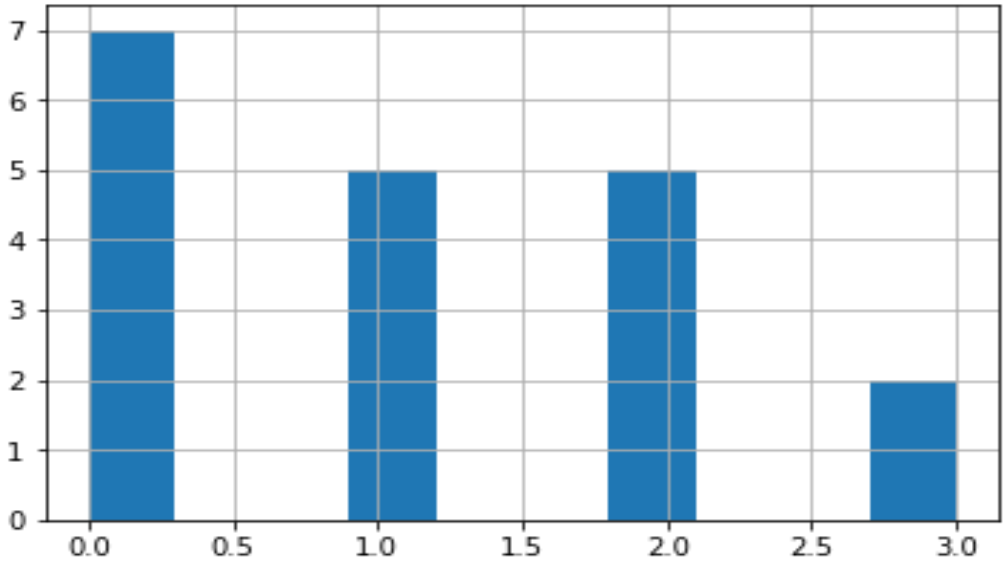


Fig. 3: Histogram for Reversed Cases

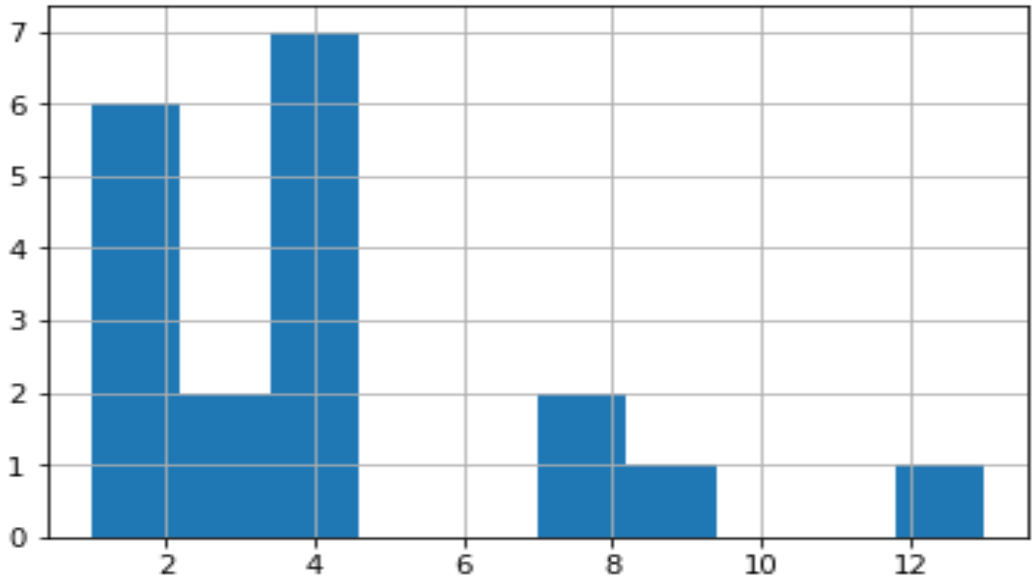


Fig. 4: Histogram for Affirmed Cases

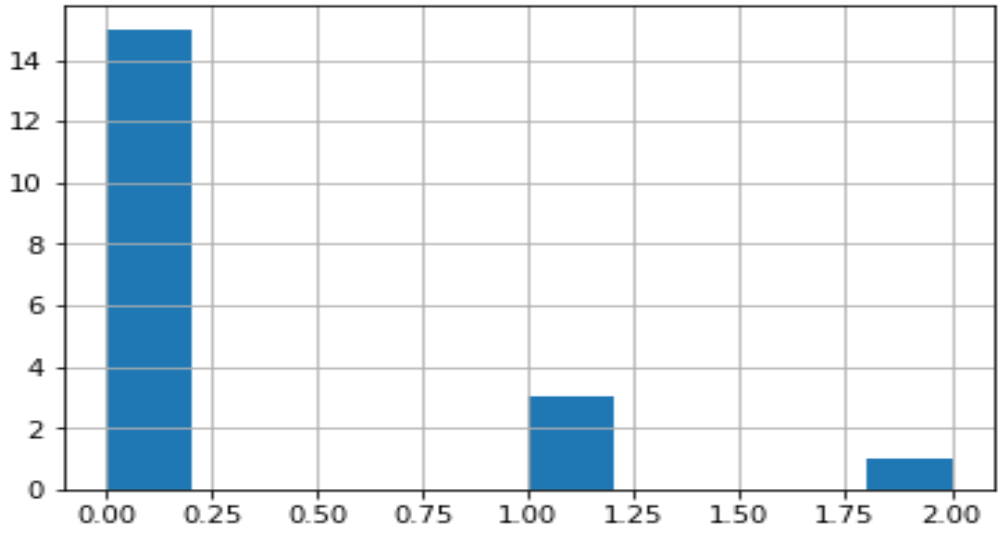


Fig. 5: Histogram for Partially Reversed Cases

Now, for the time being, we shift our focus to the cases pertaining to the period of 2008 - 2018. As done before, for this list also we are sorting the names of the respective judges based on descending order of the number of cases administered by them. This data contains only civil cases. We cleaned the data and got the data frame like below(in descending order of total number cases, 387 rows in total)

Along with this, we are going to show statistical comparisons of the decisions of the respective judges with respect to the two given periods (2018 - 2019 & 2008 - 2018). One way of demonstrating this is through the construction of pie-charts. Typically there are six pie-charts constructed using the data. The corresponding figures are shown as follows:

	Reversed	Affirmed	Partial	Total
Dennis J. Curran, J.	5	86	3	94
Alexander H. Sands, J.	5	57	5	67
Patrick F. Brady, J.	5	57	5	67
Robert A. Cornetta, J.	8	53	3	64
Thomas R. Murtagh, J.	4	52	6	62
Christopher J. Muse, J.	6	50	4	60
Linda E. Giles, J.	10	47	3	60
Geraldine S. Hines, J.	5	50	2	57
Bonnie H. MacLeod-Mancuso, J.	4	50	2	56
Keith C. Long, J.	2	50	0	52

Fig. 6: Table depicting a sorted list of the number of cases pertaining to specific judges during the period of 2008 - 2018.

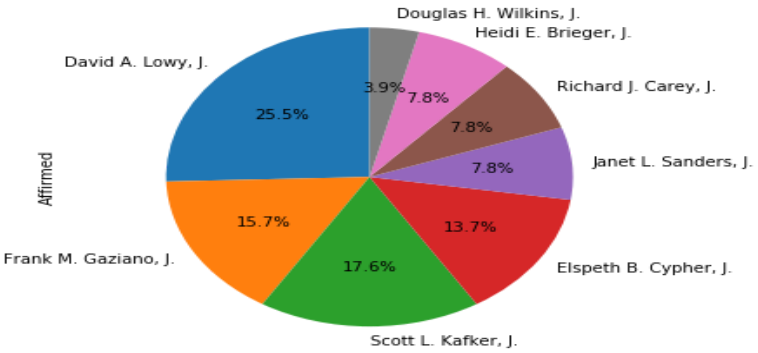


Fig. 7: Pie-chart depicting the percentage of cases affirmed by certain judges during the period of 2018 - 2019.

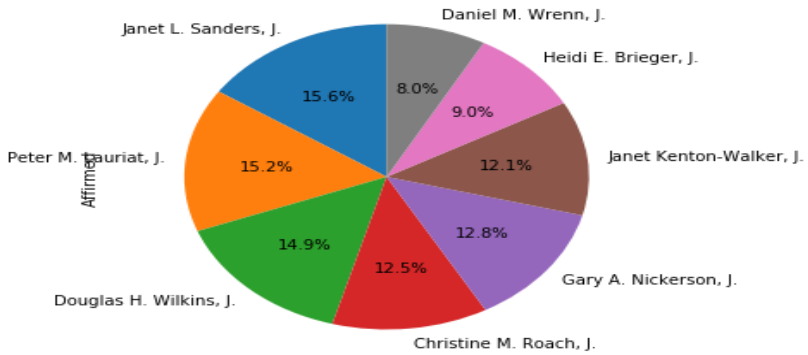


Fig. 8: Pie-chart depicting the percentage of cases affirmed by certain judges during the period of 2008 - 2018.

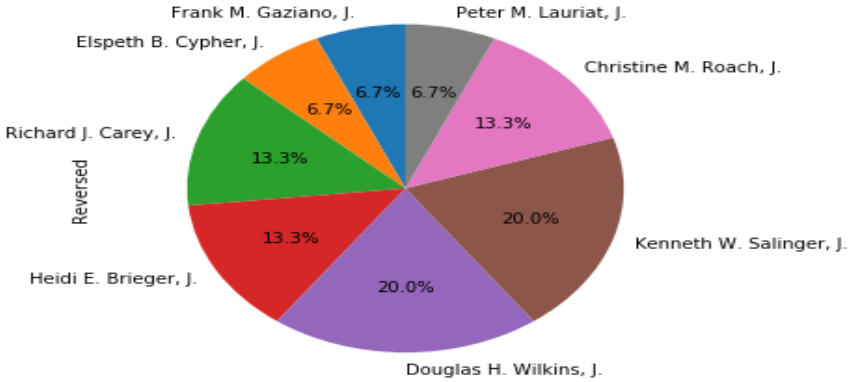


Fig. 9: Pie-chart depicting the percentage of cases reversed by certain judges during the period of 2018 - 2019.

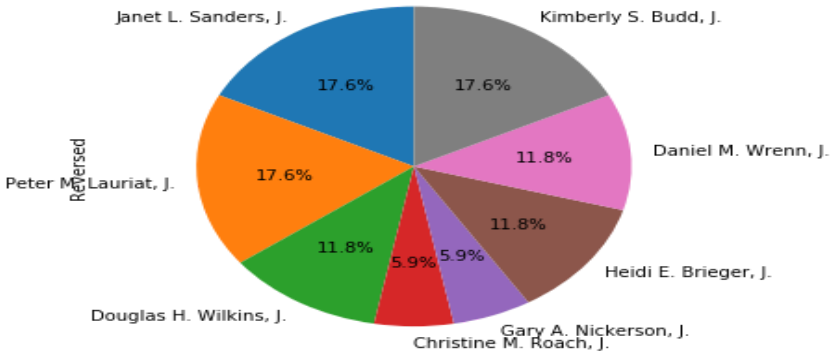


Fig. 10: Pie-chart depicting the percentage of cases reversed by certain judges during the period of 2008 - 2018.

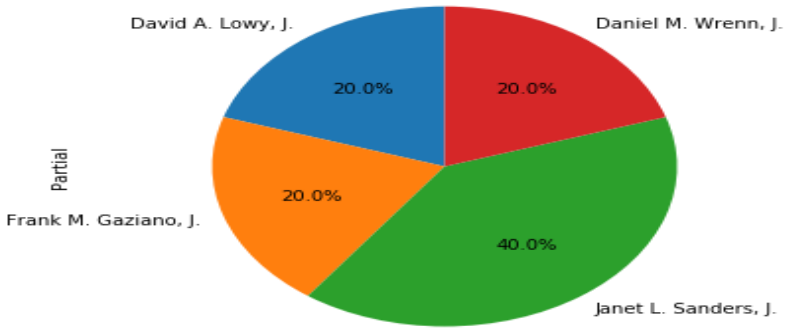


Fig. 11: Pie-chart depicting the percentage of cases partially - reversed by certain judges during the period of 2018 - 2019.

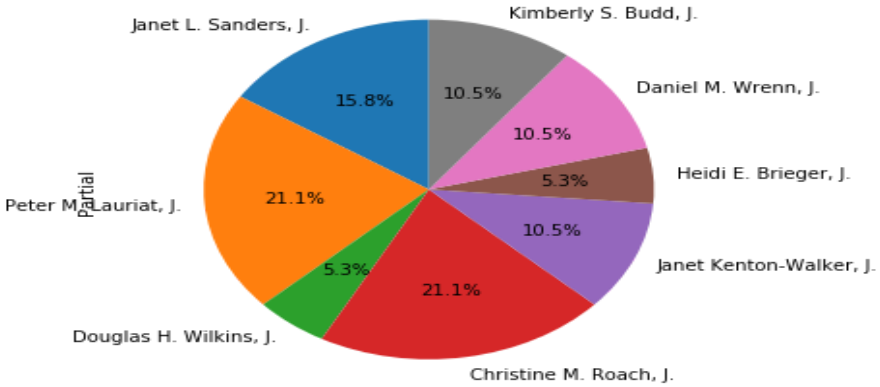
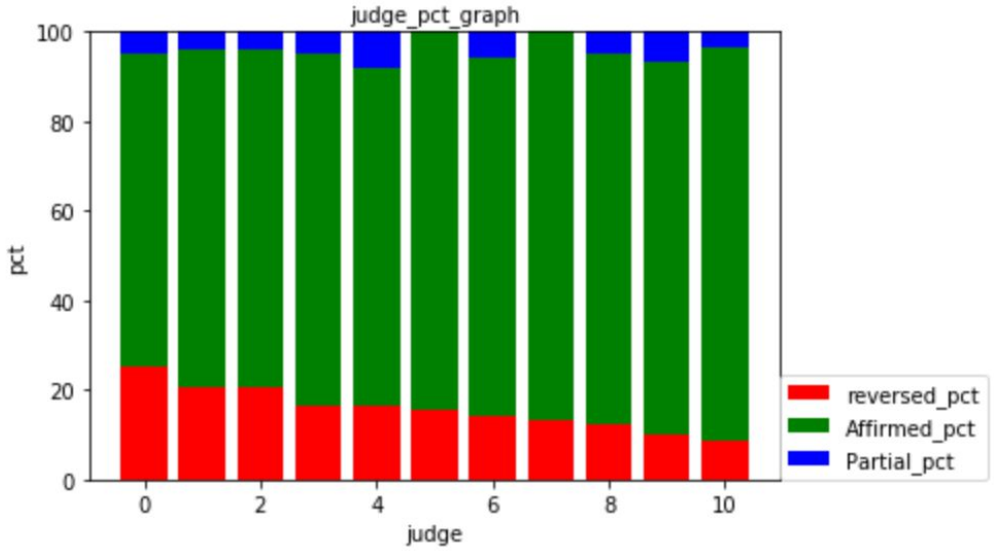


Fig. 12: Pie-chart depicting the percentage of cases partially - reversed by certain judges during the period of 2008 - 2018.

In order to find the judges whose decisions were reversed the most, we needed to explore more. We need to not only consider the absolute value of the reversed cases but also the percentage of the reversed cases for a judge. In order to accomplish this, we first find all the judges who have the total number cases larger than 12 (the average over judges' total), and reversed cases larger than 4; then we find all the judges who have the reversed percent larger than 8%; finally, we combine both results. The data frame below includes information for those specific judges:

	pct_reversed%	pct_affirmed%	pct_partial%	Total
Virginia M. Ward, J.	25.00	70.00	5.00	20.0
Elizabeth M. Fahey, J.	20.83	75.00	4.17	48.0
Timothy F. Sullivan, J.	20.83	75.00	4.17	24.0
Linda E. Giles, J.	16.67	78.33	5.00	60.0
Robert J. Kane, J.	16.67	75.00	8.33	36.0
Garry V. Inge, J.	15.62	84.38	0.00	32.0
John C. Cratsley, J.	14.00	80.00	6.00	50.0
Robert C. Cosgrove, J.	13.16	86.84	0.00	38.0
Robert A. Cornetta, J.	12.50	82.81	4.69	64.0
Christopher J. Muse, J.	10.00	83.33	6.67	60.0
Geraldine S. Hines, J.	8.77	87.72	3.51	57.0

Fig. 13: Data-frame showing judges with the cases with the most reversals.




```
Judge Virginia M. Ward, J. is represented by number 0.
Judge Elizabeth M. Fahey, J. is represented by number 1.
Judge Timothy F. Sullivan, J. is represented by number 2.
Judge Linda E. Giles, J. is represented by number 3.
Judge Robert J. Kane, J. is represented by number 4.
Judge Garry V. Inge, J. is represented by number 5.
Judge John C. Cratsley, J. is represented by number 6.
Judge Robert C. Cosgrove, J. is represented by number 7.
Judge Robert A. Cornetta, J. is represented by number 8.
Judge Christopher J. Muse, J. is represented by number 9.
Judge Geraldine S. Hines, J. is represented by number 10.
```

Fig. 14: Visualisation showing names of specific judges with the proportion of cases belonging to each judgment category

4. METHODOLOGY

- **DATA EXTRACTION**
  - Most of the data were extracted through the process of web-scraping.
  - Beautiful Soup is a Python-based package for parsing HTML documents. We exploited this functionality to get the desired information
  - The initial codebase was borrowed from the one used by the previous team.
  - Owing to the structural changes that occurred in the respective web-pages over a period of time, it was very difficult to extract the specified information from only one website.
  - For gathering the core information, two different websites were taken into consideration, namely - http://masscases.com/ and https://www.ma-appellatecourts.org/.
  - As mentioned initially, we had to scrape a large amount of data manually.
  - The scraped data for the cases pertaining to the duration of 2018 - 2019 was stored into a JSON dump.

- **FEATURE ENGINEERING**
  - For analyzing the civil cases (2008 - 2018), the dataset had about 149 columns, most of which, were not essential for gathering crucial insights. Hence, the technique of dimensionality reduction was employed to remove the non-essential columns.
  - One other feature engineering technique we employed was that of dataframe-combination. We applied this owing to the fact that we could exploit a larger amount of data when we combined the civil cases pertaining to 2018 - 2019 and that of 2008 - 2018.

- Encoding categorical variables into numeric features was necessary for generating statistical visualizations and the application of probable machine learning algorithms.
- We also removed rows that had NaN values for both the columns (‘Has Affrim’ & ‘Has Reverse’) pertaining to the dataset of cases of 2008 - 2018.

- **MACHINE LEARNING TECHNIQUES**
  - For the set of new cases from 2018-19, we tried to see if the text of the cases could be a predictor of whether a case would be reversed or not.
  - We used the scikit-learn package in python which provides libraries for implementing various machine learning techniques.
  - We cleaned up the text data using pre-processing techniques such as stemming, removal of stop words, etc.
  - We trained a LinearSVC model on this data and tried to predict if the cases would be reversed, affirmed or partially reversed.
  - On the training data, our prediction accuracy was 100% and on the testing data, our prediction was 67%.
  - This implies that our model was severely overfitting. This probably has 2 remedies: (a) More feature engineering is required to extract relevant ones from the text. (b) More training data is required.

5. CONCLUSION

As part of the analysis, we were able to generate decision reversal percentages of certain judges. However, this rate of reversal can vary significantly as we analyze data that is relevant during the period of the 1990s. The previous team was successful to an extent in discovering the similarity conditions between different criminal cases. In our work, where we have done most of the work on civil cases, there have been no clear cut associations that can identify similarities in judgments relating to the cases.

We have applied data mining and machine learning algorithms for predicting the judgments based on the data we have. As mentioned before, due to the lack of large-scale data, it is not advisable to use these algorithms on new cases. Due to the small-scale training data, the algorithm might give erroneous results when tested on new data.

6. FUTURE WORK

To get better analysis and results, several things need to be focussed on. First of all, more data would be required because machine learning and statistical models require a lot of data to provide generally applicable results. This would require the use of advanced scraping techniques since the court websites we used have high security. Secondly, legal expertise would be required to extract relevant features from the data obtained. A third option that can be recommended is to have a uniform digital format of court judgments available on the respective website. Legal academia is something that is very difficult for machine-learning algorithms and text-processing engines to understand given the complex and lengthy matter present in the documents. Human intervention is required to input judgments as affirmed or reversed based on understanding. For this, a large amount of data is required for training algorithmic models. It would also be helpful if we get access to unpublished opinions. Court judgments and related information are sensitive and hence, it is difficult to scrape data. Usage of VPN is needed for speedy retrieval.