Theme of project:
Understanding the impact and analyzing the trends after the CPDB went public. Our project compares and analyzes the trends before release of CPDB (before 2015) and compares it with the trends observed after release of CPDB (from 2015 onwards)**.**

In this checkpoint, we will be exploring whether the distribution of data from 2015 onwards different from the data before 2015.

Machine Learning:

Q1. **In the first question, we answer the question about the data distribution by creating an ML model which flags officers as "bad cops" or "good cops". This is done by training the model on data before 2015 and then testing it on data from 2015 onwards.**

We hypothesize that CPDB has changed the data distribution by introducing positive trends from 2015 onwards. Hence, the testing distribution should be different **leading to low accuracy on the test set.**

We start by creating the relevant data. We have created CSV files having information on good cops and bad cops. The cops which do not appear in the data_officerallegation table are classified as good cops. Refer to ML_flagging_1.sql for the data creation code. (Under Extra_Local_src_files\CP5\code)

Five features were used for the classification task: Race, Rank, Age, Appointed Date (Year only), Number of Awards

We trained the dataset on two different classifiers:
Here train accuracy is accuracy on data before 2015 and test accuracy is accuracy on data from 2015 onwards.
Results on Logistic Regression:
Train accuracy is: 0.6667755635413264
Test accuracy is: 0.13287398651658697

Results on SVM:
Train accuracy is: 0.6631819666775564
Test accuracy is: 0.13055745300109886

Analysis: As we hypothesized before, the test accuracy is significantly low as compared to the train accuracy. This is a strong indication that the test distribution is different from the train distribution and release of CPDB has had an impact on the trends.

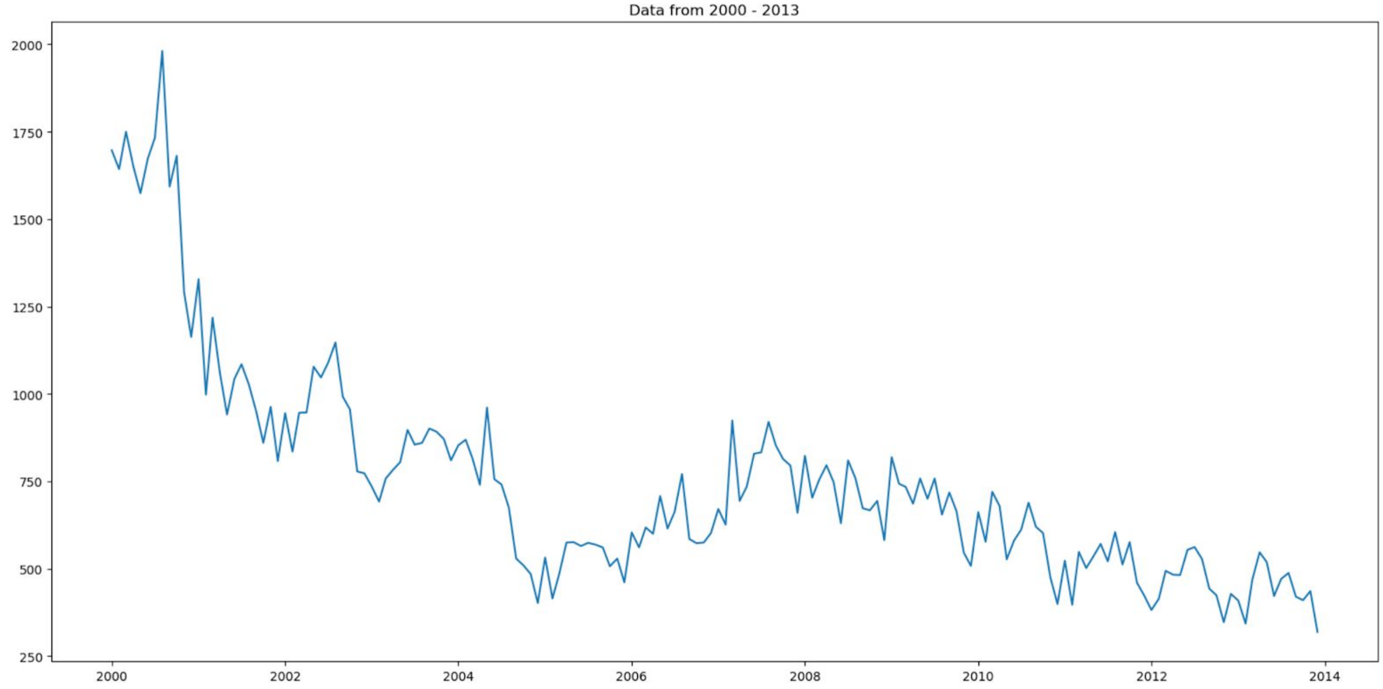
Interactive Google Colab notebook with code: (Please refer to the README for execution instructions)

https://colab.research.google.com/drive/13KmfroDv8ZJXdK4Limb4vM3OCh3G3rXB

Q2. In the second question, we created an ML model which can predict the number of complaints in upcoming years. We trained the model on data before 2015 and then tested it on data from 2015 onwards.

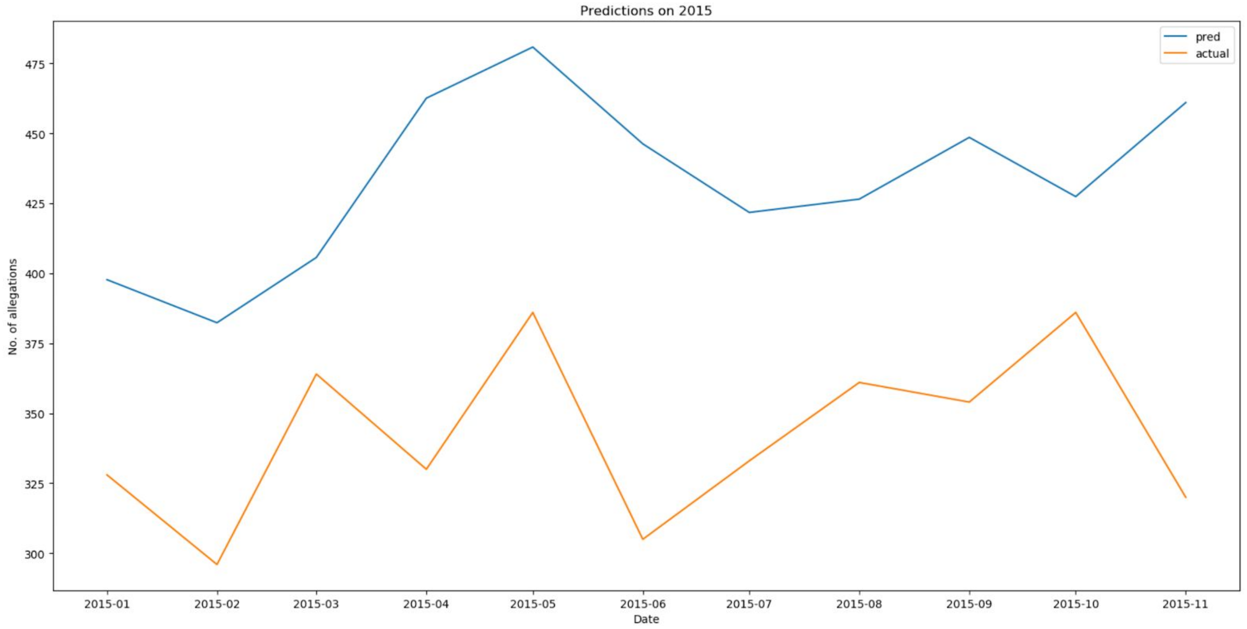
Since our aim is to see how the no. of allegations have changed over time, we made use of time series analysis. For this question, we performed time series analysis using the ARIMA (Auto Regressive Integrated Moving Average) model.

First, we visualize the data distribution before 2015 as follows:



At this point we know that the number of allegations have gone down after 2015 since the cpdb went public. The question that comes to our mind are: What would have happened if cpdb had not gone public?

Our model tried to predict the no. of allegations for the year 2015 assuming that the data is following the same trend as it did over the years. Thus, the model gives predictions for number of allegations in 2015 if the cpdb had not gone public. We compare this result (cpdb did not go public) with the actual result of 2015 (cpdb went public) to see how much of a difference there is.

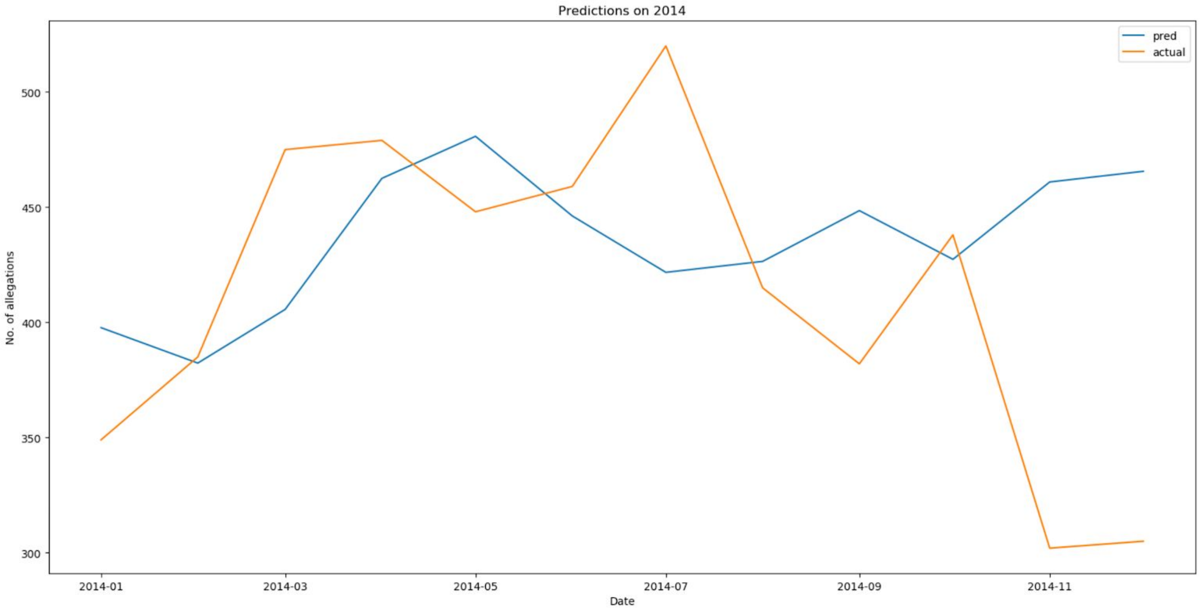


We got a high difference between the predicted and actual values which confirms that cpdb had a positive impact on the no. of allegations. However, these results could be because of two possibilities:

Possibility 1: The model performed poorly. In this case the results that we got will not make any sense since the model was bad and hence we got the difference between the predicted and actual values.

Possibility 2: The model did a good job predicting the values, which implies that our results that the predicted and actual values are very different is correct

So we performed one more test. We created a test set for the year 2014 (at this time cpdb had not gone public) and compared the values with the values predicted by our model.



Our model performed well on this data and hence we could conclude on Fact 2.

As we can see, the model does a pretty good job of following the distribution of 2014.

Using another approach, we took the following **features** for each officer: race, gender, rank, birth year, salary and number of awards, and train **two models** using it- Linear and SVM regression.

Again over here we train on data before 2014 and then test on data for 2014 and 2015.

Results:

```
LR R2 score before 2015 -2.419732473818787
LR R2 score after 2015 -5.691996237628087
SVR R2 score before 2015 -1.18523807088363
SVR R2 score after 2015 -1.9501727441939156
```

So, again our hypothesis is validated that from 2015 onwards there have been significant changes in the trends.

Link to code:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3088900325675746/4281434282507131/3349831233202429/latest.html

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3088900325675746/4484316082940702/3349831233202429/latest.html

Q3. For this question, we group similar words in different categories from document text using the N-gram model and then try to analyze the differences in the groupings before 2015 and from 2015 onwards.

In this, we found out that due to the sparse nature of true labels in each category and the extremely noisy data (misspelled names, mislabelled data), the resulting groupings weren't great. Hence, it's difficult to perform analysis on the nature of distribution with the data available to us.

Still, we found some interesting trends.

Let's consider the category of nudity and penetration.

Before 2015: Words like 'clothes', 'pants', 'removed' were found. The word 'stomped' was strongly associated with the word 'removed' which kind of shows the general aggressive nature of the officers before 2015. There were also words like 'restaurant' and 'street' which gives us an idea of the kind of places were incidents of these category took place.

After 2015: Words like 'parties', 'vehicles' were found which shows the kind of places incidents of these categories took place.

Another interesting observation is that words like 'drugs', 'cannabis', 'narcotics' were present in groupings from both before 2015 and 2015 onwards showing that incidents of this category strongly overlap with incidents of drugs.

Such trends can be studied for the different categories by observing the probability distributions given by our code. But again we would like to reiterate that the results aren't conclusive as the data is too sparse.

Other interesting observations:

- Officers with rank lieutenants seem to be strongly associated with the category tasers
- Officers with rank detectives seem to be associated with the categories - Trespass, Tasers, Racial Slurs and Neglect of duty.

Link to Google Colab notebook: (Please read README for execution instructions)
https://colab.research.google.com/drive/1Ck1PhS12NVUhW_RnHihbONFcKILQxatq

Q4. Developing a model to give a “severity measure” to each document

Here, we try to assign a numeric severity value to each of the summaries present in the dataset. Majority of our efforts went in analysing the 1800 rows long document to get relevant words, phrases and patterns and assigning them a weight to contribute to the overall severity value. We have assigned these weightages based on the average number of years of punishment that is needed to be served for that crime. To make sure multiple lesser severe incidents don't get a score greater than single, but much worse severe incident, we have kept the weights in powers of two. The weights were assigned based on 10 levels of crimes. We use fuzzy pattern matching with a threshold of 0.9 for matching the words.

Results:
After assigning a numeric severity score to each of the documents in the dataset, we analyse some of them. The highest severity score that any document got was 275. This particular document had matches for the following words phrases and patterns:
[['raped her', 'tried to murder', 'drugs', 'choked', 'robbery', 'his clothes', 'operation/personnel violations']]

This model can be extended and used to compare the average severity before 2015 and after 2015.

Link to notebook:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3088900325675746/846831505272150/3349831233202429/latest.html