Deliverable 2:

For this deliverable, we completed all of the to-dos on our Kanban board. We started by adding the "Allegations" and "Findings" columns from the Boston Globe disciplinary action database to our dataset. Doing so allowed us to have more context of the disciplinary action and punishments, which can later be used for further analysis. Then, we completed step two of our suggested steps (refer back to Deliverable 0) which asked us to merge the disciplinary action database with the BPD financial contributions data and verify common names using the Race/Ethnicity BPD personnel dataset. The process we took to complete step two is described below:

Preprocessing
For preprocessing, we had to prepare the two main datasets to have names that were formatted the same in order to allow us to better merge them with fuzzy matching. We created a python notebook named "preprocessing.ipynb", which preprocessed the names. For the disciplinary action dataset, it was not as complicated as names were *first name middle name last name suffix*. However, some names were capitalized and others not so to preprocess this column we simply put the entire name in lowercase. For the contributions dataset, it was more complicated. It is formatted as *last name, suffix, first name middle name*. So, we transformed it into a list separated by "," and depending on how long the list was we rearranged the names in order to make it formatted the same as the disciplinary action dataset. We additionally made these names lowercase as well to match the disciplinary action dataset.

Fuzzy Matching
For fuzzy matching, we used the "fuzzy matching template.ipynb" provided by our PM, Gowtham. The code first splits each of the data frames by the first character of the last name. We then wrote a function called getLastCh(s) to create a column with the first character of the last name. Then, we merged two data frames using their lastName characters and applied a string similarity score. For each row, we filtered the string similarity value to create the final dataframe with name matches. This merged subset is then written to a CSV and then we repeat this for all last names that start with each character of the alphabet. We then merge all these subsets back together for the final merge.

Finally, we filtered the merged dataset for people that had listed "Boston Police" as their employer. This final merged and filtered dataset can be found here.

Next Steps
Now that we have our merged dataset, we can move on to the next step which includes first overlaying the LEAD police blacklist officers onto the merged dataset which should add another 54 BPD officers. After that, we will compute the campaign contribution totals between all BPD

officers, as well as only officers under investigation. This will allow us to then analyze and visualize the data to find any patterns between officers under investigation and political contributions.