MSAI 339 project proposal 2nd version

research proposal:

We want to explore the relations among misconduct settlement amount, officers' overall background(age, race, rank, salary, status, gender, award, promotion, etc), victim overall background and the details of misconduct(time, weather, location, and location's detail, type of misconduct, weapon).

Someone may want to predict future settlement amount for each police officer, but we do not think it is a good idea, because "bad cops" are only a small portion. This means most settlement amount data of police officers are 0, and they have no misconduct history. This is as hard as goods suggestion for customers without their previous shopping record.

Therefore, we want to train a model to help predict every future misconduct settlement amount using its related officers and victims' background and details of misconduct. This is a more fine-grained prediction than predicting the number of misconduct cases and settlement amount of them in every district or block.

research questions:

1. relational analytics
    a. what's the relation between average number of complaints received per year and the age of a police officer? older ->more complaints or something else?
    b. Do those who receive no complaints get more awards/promotions?
    c. what's the average number of complaints received per year and officer?
    d. what's the average number of complaints from different racial officers per year?
    e. what's the average income of those officers who have received complaints and the average salary of the whole Chicago police officers?
2. visualization
    a. Does racial discrimination really exist in misconduct cases?(bar charts will be applied, x-axis is different races, y-axis is the number of complaints from each race divided by the population of corresponding race)
    b. which district/part of Chicago city cost a higher amount of settlement(bar charts, x-axis is different districts, y-axis is the amount of settlement at each district)
    c. Are men and women treated equally by police officers? Will women report more misconduct cases than men?(bar chart, x-axis is two genders. y-axis is number of complaints from each gender divided by the corresponding population)

    d. how is the number of misconduct cases and their total settlement amount at Chicago changing?
3. data cleaning and integration
    a. what's the possibility of those who received more than 5 complaints or less than 3 complaints commit misconduct and total settlement amount? are those who received more than 5 complaints commit worse misconduct and cause more settlement amount?
    b. what's the average possibility of misconduct and corresponding settlement amount for different ranks(supervisor, investigator, detective, police officer) of polices?
    c. what percentage of officers who are named in settlements have **never** received complaints?
    d. what percentage of officers who are named in settlements have received awards/promotion?
4. graph analytics
    a. can we identify the most likely to be co-accused cluster of police officers, retrain them and split them out to make them never co-work in an unit again?
    b. what's the average settlement amount per case and what's the average number of misconduct per officer in each cluster?
    c. for those who have co-worked with the top1% most accused police officers. what's their previous average number of complaints before he worked with the "very bad" cops and after? Does the number increase significantly?
5. machine learning and text analytics
    a. how can we predict each case's settlement amount using its related officers and victims background? what fields are important? like race, awards, income?
    b. what machine learning models can make our future misconduct settlement amount prediction more robust and accurate with relatively less training and inference time?
    c. what's the average settlement amount for misconducts with different tactical response tags?
    d. what's the number of misconducts with different tactical response tags, which ones are more frequent and should be paid attention to?