Proposal of CS506 Coursework

---BPD Payroll Investigation

Anzhe Meng, U50590533
Ruizhi Jiang, U17637349
Jiahao Song, U57363411

Our team is assigned to collaborate with BU Spark! and Mr. Paul Singer wants to dig into the police's payroll and find out the potentially ill-behaved but not decertified police officers. Now in the rest of our proposal, we are going to present to you what we plan to achieve and how we are going to achieve by the end of this semester.

- **Why do we do this?**

According to Mr. Singer, nowadays in Massachusetts there exist some policemen or policewomen whose wages can hardly match their contribution to our community. This phenomenon is, in general, due to two reasons. First of all, some ill-behaved police officers are relocated after they are decertified in one certain city. Given their past un-disciplined records, probably they would behave as badly as before while being paid as usual. Furthermore, some policers ask for overtime or injury compensation though it is not necessary at all.

No matter for which reason, the existence of this kind of act is definitely a waste of tax and public resources. In order to expose this situation, Mr. Singer expects us to discover the pattern from the existing datasets as the proof for their journalistic investigation. For example, we are expected to find the policemen/ policewomen who were decertified while still being paid. Generally speaking, our goal is to help improve the service of the police force so as to contribute to building a better community.

- **Goals, hypothesis and outcomes**

Generally speaking, our goal is to find the decertified policemen who are working in the new places. Given their unknown undisciplined record in the previous place, they are working in the new places as the qualified police officers. But we should be aware of and minimize their potential risks to the society, though they may behave like disciplined policemen.

For the part of hypothesis, first, we assume the data we collected are sufficiently correct and valid. This is the foundation of further data analysis. Second, the different data we collected can be possibly merged. This is also important, because if we want to conduct further data analysis towards the diverse data sources, we need to compile and merge it at the first stage. For example, suppose one data table includes only the names of officers

but their badge codes while another data table only show their codes but names. In this case, it is impossible for us to find out their own relations simply according to their column names. Fortunately, our present datasets satisfy our assumption, facilitating us to dig deeper into them.

Finally, we will be able to answer the questions related to the payroll data. For example, we could get the officers who were paid most in Boston. We could also know what kind of officers were paid the most in terms of overtime/injury. These are the questions required by Mr. Singer. Once we could answer some basic questions, if time allows, we will explore more patterns in the data, flag promising leads for possible further investigation.

- **Non-goals, out of scope topics**

After we finish what the project requires most urgently, we are still willing to explore something out of scope topics. For instance, we could delve into the topic of making BPD more efficient based on the data we collected, which is beneficial for Boston residents.

If our project progresses well, based on our research, we could also provide suggestions to governments, telling them how to supervise the group of policemen/ policewomen and how to make policemen disciplined, even

telling them how to identify a policeman/ policewoman with decertified record in other places.

- **End result/product**

By the end of the project, in theory, a payroll-oriented software (or program) would be designed and implemented for WGBH. To be specific, we expect our program will first help merge several data sources provided by WGBH into one dataset. Then, by extracting valuable features from the dataset and analyzing them, the software would automatically provide us some basic information of officers in MA and answer the questions given in the program description. Most importantly, it would also figure out who are the ill-behaved police officers we are trying to find and where they live now after being decertified. Finally, results analyzed by the software would be shown in the form of graphs.

- **Open questions, uncertainties**

Apart from the questions we raised above, there are also some open questions we would consider. Firstly, when learning the pattern of ill-behaved policers, can the software also provide extra information for us? For instance, can it analyze the geography information, such as the general pattern of the living neighborhoods, of their previous living place? This may influence the accuracy of predicting whether the policeman is ill-behaved.

As for uncertainties, right now we could only name two because what is uncertain is also unpredictable. First, we may scrape the data from some websites. We do not know those websites are equipped with the scraper-detective mechanism. If they are, chances are that it’s a little bit challenging for us to gain what we need.

Second, after we get the data we want, like it is said above, we need to merge those data so as to compare them and make them meaningful for us to analyze the data. So, we recognize there would be possibly some format issues awaiting us. Merging the data smoothly and efficiently is an important step in our whole project life cycle.