Midterm Project Proposal

Keliang Xu

11/4/2021

## Personal statement

After half a semester of study, I thought about a lot of work I would like to pursue in the future. I hope to be a data analyst and do data consulting in large technology companies. My thoughts on my career are still rather vague. But I think as long as I am proficient in data analysis skills, it will definitely help my career. I hope to be able to grasp the ability to properly resolve any aspect of the data that sent to me. This requires me to have the ability to quickly learn and understand other aspects in addition to the ability of data analysis. Therefore, I want to choose a data content that I have relatively untouched, and complete a question that is challenging for me.

The database on the Kaggle website can be seen as a preview of future work. In my vision, I hope I can properly handle any given data. Because this website is also raising questions from some companies, I think the task of processing data in this website is close to the content of working in the company in the future.

Regarding crime data, it has something to do with my own interests. I think that some meaningful conclusions can be drawn from this data. For example, show the places and times with high crime rates on the map, and strengthen police patrols in this place on the government side; at the company level, you can design emergency alarm software to help and improve people’s vigilance; at the personal level, people can also know the safety of their location.

## Question

There are three questions that come with Kaggle data, and three questions that I want to study.

-What types of crimes are most common?

-Where are different types of crimes most likely to occur?

-Does the frequency of crimes change over the day? Week? Year?

-Boston crime rate map (at different time)

-What factors are related to crime frequency？ (time, place)

-What factors are related to the type of crime (time, place)

## The data source

The data set is Crimes in Boston, including times, locations, and descriptions of crimes.

<https://www.kaggle.com/AnalyzeBoston/crimes-in-boston?select=crime.csv>

Containing 17 columns and about 320,000 rows of data, I think this is a relatively large data set I have come into contact with. At the same time, this data is more in line with my ability to process data now.

## Timeline of work

If the time plan is not strictly implemented, it is easy to postpone the work later. Therefore, I hope to be able to complete the task strictly according to the time plan, with a maximum of one day’s grace for each due.

1.Data cleaning and simple data analysis: 11.4-11.7

2.EDA: 11.8-11.12

3.Modeling and Validation: 11.13-11.23

4.Write up: 11.27-11.30

5.Polish and submit: 12.1-12.5