MA678 Midterm project proposal

Name: Lin Zhou

Personal Statement: I am interested in data related to nature and artificial environments where we are living in like the weather, geography, air condition or road, building design and so on. From my perspective, the analysis could give people suggestions and guidelines which might save people’s lives and provide a better living place for them. Since I am a beginner of data analysis, it would be helpful for me to practice more about EDA and following process. I look over different resource online and find the US traffic accidents records dataset is suggested to be suitable for beginners to do some data analysis practice. These data were derived from real life. As a result, from my perspective, they are easy to understand and after data analysis, the results could be interpretable and might lead to some practical directions. I think this practice can help me with getting familiar with the whole process and solve the real problems. I am also interested in making predictions after the data analysis. In this case, the time and the place of potential traffic accidents might be predicted if combined with the weather forecast or other information. Then precaution like warning board, speed bumps could be set to alert related personnel, which would decrease the numbers or possibility of traffic accidents. It is a very practical dataset, I think I can apply the knowledge like data cleaning, different models learnt from textbooks to the reality. The difficulties and challenge I will meet during the project would give me some guidelines which might solve the problems in the real work in the future.

Possible Questions: Which factors would affect the severity of a traffic accident? Where and when would the traffic accidents most possibly happen? There are many aspects the dataset can show us during the data analysis process like the overall summary of whole country traffic accidents, comparison of traffic accidents among each state.

The data source: US traffic accidents records (2016 – 2019)

<https://www.kaggle.com/sobhanmoosavi/us-accidents>

The dataset contains details of 3.5 million traffic accidents that took place in the United States, from February 2016 to June 2020. There are total 49 columns inside the file including severity, accidents start time and end time (in local time zone), place, weather, temperature and moisture so on. I would like to fit the data into different models to find the suitable model

Proposed Timeline of work:

EDA: 15th Nov

Data Processing: 20th Nov

Modeling and Validation: 27th Nov

Write up: 2nd Dec