Project Proposal – Data Visualization

1) **The dataset**

The dataset is divided in 4 different parts, from the year 2001 to the year 2017. The total of entries across all the different parts is around 6 million and it has a total of 22 features.

Exploring the dataset, it is possible to see that some features have a considerable amount of missing values, for example the features related to geographic location.

Since the number of features is quite large, it can slow down the pre-processing a bit, so for each problem we single out the most relevant features. For example, when the problem requires geographic location we only consider features that provide this location and the type of crime/year to answer the question proposed.

2) **The questions you expect to answer by analyzing the data**

What are the safest areas in Chicago? How did safety levels develop with time?

Are the efforts in mitigating crime being effective? Do they contribute to a safer society? Does crime follow any temporal pattern?

Which beats are overloaded? Where should the police focus when dispatching cars?

3) **The hypotheses you formulate by an earlier inspection and exploratory analysis of the data**

We believe crime has decreased in general terms, but there might be differences across segments. We believe the police might not be addressing crime types in proportion to their volume/seriousness.

We believe there might be some problems related to police staffing, making them less reactive to simultaneous crimes in the same location.

4) **The problems you foresee in analyzing the data and creating visualizations**

Modelling the safety function and deciding what weights to use. Effectively sampling the data. Transmit safety levels for a given year (present) and simultaneously represent their time evolution in an effective and simple way. Scaling seriousness of crimes. Deciding on resolution vs simplicity in maps.