**Challenges 2: Help the Chicago Police Prevent Crime!**

Group name: the bootsrappers

Group members: Einat Sarig, Niv Ben-Salmon, Keren Ben-arie, Dana Adam

STEP 1- Cleaning and Preprocessing of the data

We decided to divide the data by 70% train, 20% validation and 10% test.

First, we made some basic validation checks, making sure we are getting the proper types in each feature and that the given values are legal.

After that, we tried to determine which of the features is relevant to our analyzing process.

We realized there are many features describing the location and started off with a correlation map between them:

Square

Description automatically generated

As we can see in the attached plot, there is a really high correlation (almost identical) between x coordinate and longitude, the y coordinate and latitude, beat and district. In that case, there is no need in holding all of those, so we decided to delete the longitude, latitude and district features.

In addition to that, we removed the ID, Case number and Year features because those had no influence on the decision process (all crimes given had the same year).

The date feature was a bit more complicated since it has date and time- 2 different features that need to be considered separately since they might have different influence on our predictions, so we split it to 2 different columns.

For the final step of pre-processing, we analyzed the free text column - Location description. We chose 20 different significant keywords and made a binary column for each one of them in order to identify a certain behavior based on specific sites such as schools, banks, streets etc.

STEP 2- Building models and training