Data collection

To examine this question, we use open data about crime and policing in England, Wales and Northern Ireland, delivered by the UK police on data.police.uk.. The UK police delivers API’s containing monthly data about every stop and searchers for each police force in England, Wales and Northern Ireland from July 2016 to May 2019. Each API contains data for all stops and searches for a given month in each police force. In order to gather the data from each police force in each periods into one DataFrame, we had to loop through every month from July 2016 to may 2019 (the inner loop), and then loop through every 44 police forces in England, Wales and Northern Ireland (the outer loop). In each inner loop we created a new column variable *force*, which indicate in which Police force the stop and search took place.

In total we scraped XX API’s each containing 17 columns, with a total of 838034 number of rows. The included columns is …..

The police API have a rate limit of 15 requests per second with a burst of 30, which meant that we could easily set the waiting time down to 0,1 seconds (kilde: https://data.police.uk/docs/api-call-limits/). We also had to put the timeout up to 180, since the scraping of the API for the London police force didn’t succeed in first try, duo to to many observations. This can be seen in the two figures below, showing respectively the call time and the response size over time, where there is a spike when calling/scraping the London police force. In the first figure showing the time it took to make the call, and on the y axis the additional time each request took. Around 09:35 there is a huge spike, showing that the call time for one specific force, took

Et billede, der indeholder skærmbillede

Automatisk genereret beskrivelse

Et billede, der indeholder skærmbillede

Automatisk genereret beskrivelse

Data cleaning and transformation

Many of the observations for stop and searches had missing values. This is not a result of an error in the scraping process, but simply because the police have not filled in all information for al the stops and searches, hence it was possible to reestablish the missing information. Therefore, we had to drop all rows where there missing values for the variables included in our machine learning model. Neither the British Transport Police nor the Police Service of Northern Ireland provides outcome data, hence data for both forces is dropped, hence we have no information from Northern Ireland in our analysis. Under the assumption that there is systematic in which stops and searches that has missing values, dropping them will not influence the usefulness of the machine learning model. This assumption can off cause be discussed, with we will get back to in the XX part.

In order to improve the prediction of our machine learning model, we wanted to create more features to include in the model. To use both the time of the day and the time of the year for the stop and searches as features, we had to subtract month and the hours from the *datetime* variable, creating two new variables *month* and *hours.* We also wanted to get a more specific geographical area then the police force as a feature. In the variable *location* there is an estimated longitude and latitude of the stop and search, which could be used to locate under which local authority each stop and search took place. This was done by importing a CSV file from XXX containing the longitude and latitude of each local authority (er punktet I midten eller?) in England, Wales and Northern Ireland. By minimizing the distance between the location of the stop and search and the local authority we could approximate under which locate authority the stop and search took place. This is only an approximation since it is not sure that the local authority located closest to the location of the stop and search is the actual authority in which the location belongs. It gives however some smaller geographical areas which can be used as features in our machine learning model*.* In order to use the longitude and latitude of the stop and searches, we first had to subtract these from the dictionary variable *location.* This meant that we also got 4 new columns *latitude, longitude, street\_id* and *street\_name*. Finally, we created a dummy variable *XXX,* indicating whether the stop was a success or not, meaning whether the police found something illegal on the searched individual. The variable is defined as a “failure” if the *outcome* equals “False” or “Nothing found – no further action” and otherwise a “success”.

Our final data set contains XX columns and XX rows.