Index

1. Client requirements

1.1. Summary

Client provided us with data about the stop and search operations of police departments in the United Kingdom.

There have been accusations in the press that the police tend to stop and search certain minorities at a higher rate than others, and that women of certain age groups and ethnicities are asked to remove articles of clothing more than the others.

We must:
1) Explore the data, report on interesting findings and investigate whether the data proves any of these claims.
2) Host a model with an API endpoint to be used by the officers to approve the stopping of a person/car.

1.2. Requirements clarifications

Email sent do Dr. Wilson:

“Hello Dr.Wilson,

I am Henrique, the specialist hired to work on this project. I am excited to start working with you and hope you find our partnership pleasurable. Feel free to contact me about any matter.
Henry sent me your email exchange and I already did a first analysis on the data and requirements of the project.
To confirm that we are on the same page and to correct possible misunderstandings, here are some questions:

1) The column ‘**Legislation’** indicates the law that was used to justify the stop and search operation and then the column ‘**Object of search’** gives the specific intention of the officer for using the law, and finally the column ‘**Outcome’** with the conclusion of the stop search operation, which can be divided into *TRUE* or *FALSE* in relation to the column ‘**Outcome linked to object of search’.**

How can the ‘**Outcome linked to object of search’** ever be TRUE when the ‘**Outcome’ is** 'A no further action disposal' ?
Some of the operations have the ‘**Object of search’**: Controlled drugs, Offensive weapons, Articles for use in theft, Stolen goods,... If the ‘**Outcome’** is that nothing is found shouldn’t the ‘**Outcome linked to object of search’** always be FALSE?

2) In order to classify the success of the operation by its '**Outcome',** here is a suggestion of the division:
SUCCESS: Arrest ; Community resolution ; Penalty Notice for Disorder ; Khat or Cannabis warning ; Summons / charged by post ; Suspect arrested ; Caution (simple or conditional) ; Offender given drugs possession warning ; Local resolution ; Suspect summonsed to court ; Article found - Detailed outcome unavailable ; Offender given penalty notice ; Offender cautioned ; Suspected psychoactive substances seized - No further action
FAILURE: A no further action disposal ; Nothing found - no further action

Is this classification ok or should some more options be considered a FAILURE? Or maybe create a third category with a more neutral connotation as a result of an '**Outcome'** that wasn't very serious like Local resolution ?

3) When we have information about '**Outcome linked to object of search'**, does it make sense to link the success of the operation to a *TRUE* result in the '**Outcome linked to object of search'**, and not the SUCCESS '**Outcome'**?
This way we would be directly evaluating the officer's *reasonable grounds for suspicion* to use '**Legislation'** and the specific '**Object of search'**.
The same in relation to the quantification of the degree of discrimation, we can quantify the bias on cases of FALSE '**Outcome linked to object of search'** if the information exists, and on FAILURE '**Outcome'** when it doesn't.

4) In relation to the API, it will receive information on all columns of the training dataset file except for the '**Outcome'** and '**Outcome linked to object of search'**, and it should suggest whether or not to stop and search by respecting the following requirements:
a) minimum of 10% likelihood in the search SUCCESS per station and per search objective
b) maximum of 5% discrepancy in the search FAILURE rate between protected classes (race, ethnicity, gender) and between genders on the solicitation for the "Removal of more than just outer clothing".
c) maximize the '**Outcome linked to object of search'** given the constraints above.
Are those percentages acceptable?

5) In the case of being '**Part of a policing operation',** the decision to stop and search is constrained to the characteristics of what is being specifically searched for. Is there still space for bias on the police decision or we shouldn't consider them in quantifying discrimination of minorities?

Don't hesitate to contact me for any clarification or new information that might be relevant to the project.
Thank you.

Best regards,
Henrique Baltazar
Data Science Consultor

Awkward Problem Solutions™. “

Additional information exchanged with Dr.Wilson:

The available Latitude Longitude, is related with the ‘station’ column?’
There will probably be some correlation, but we’d expect that it was the location where the officer was then they conducted the search. Please let us know in your report if this is not what the data suggests.

How should we deal with the gender “Other” regarding discrimination?
Please exclude it from the analysis, unless you find some very strange pattern that is worth reporting. Regarding using it in modelling please use your best judgement.

The Metropolitan station (which is very large and has lots of data in the training set) has the features Outcome linked to object of search and Removal of outer clothing without any data (always missing). Is this a known problem, and if so how should we proceed?
Thank you for bringing this to our attention, I’ve contacted the administration at the Metropolitan and asked that they fix their data entry. Please include this in your report, and do not use the Metropolitan station’s data for training your models. They will not be in the test set.

What should be considered a successful search?
A search is considered successful if the outcome is positive, **and** is related to the search.

There are records where 'Outcome' is 'A no further action disposal' but 'Outcome linked to object of search' is True.
Thank you for bringing this to our attention, it suggests that there are data entry issues. Please detail this in your report so that we can take action with the stations where this is occurring.

How should we deal with mismatches between 'self-defined ethnicity' and 'officer-defined ethnicity'?
For the purposes of model training and of analysis of potential discrimination please use the officer reported values.

Should Removal of more than just outer clothing be filled with False?
Yes, except when it’s just a vehicle search, in which case it makes no sense and should be kept as NaN. Otherwise it’s considered a data imputation error.

In the briefing we were told that "the success rate of the searches should not vary significantly between populations". Is there (A) a defined target for this variation? Also, (B) how should we define population in this case? What would be (C) an acceptable difference between police station search rates?

We would hope that (A) there would not be a discrepancy of more than 5 percentage points between population sub-groups, which would be defined as (B) a (station, ethnicity, gender) tuple, and that the discrepancy between stations (average per station) would not be larger than 10 percentage points.
We are only concerned about age when deciding about clothes removal. For now we are not interested in ages when deciding whether or not to conduct a search.

How should we measure discrimination?
Our current priority is in making sure no population is over-searched, which we are defining as having equal success rates. We know that due to correlation with economic status different groups will not have the same search rate, which is acceptable for now.

The data suggests that there are subjects with ages under 10
Thank you for bringing this to our attention. Please detail this in your report so that we can take any necessary action.

Could you clarify what it means to conduct a search only when there is more than 10% likelihood that the search will be successful?
Ideally we want a system that gives a good probability of the search being successful, and if the probability is lower than 10% then the search should not happen. Ideally that will reduce the cases where the officer should "obviously" not search. Please let us know what the best way to measure this is in your report.

Is there a minimum number of operations to consider a police station?
Actually that may be worth considering at a population sub-group level, or (station, ethnicity, gender) tuples. Please consider that if they are smaller than 30 people then we have no significance, or suggest an alternative metric based on your expertise. Again, please note that we are only concerned about age when deciding about clothes removal.

There seems to be a lot of discrepancy between stations. Is this something we should attempt to correct?
Ideally yes, we're hoping that having a unified policy will in part correct differences between stations. Having said that we're more focused in minimizing differences between sub-populations (see definition above), so please let us know in your report what a good trade-off looks like.

The data changes over time, as was mentioned in the briefing. Are there any specific dates when such changes are expected to have occurred? (changes in IT systems or reporting practices)
Again this is a fair question, but we at HQ don't actually have visibility over the individual reporting systems at the different stations. If there are any large changes on particular dates we would be very interested in learning about it in your report.

The available Latitude Longitude, is related with the 'station' column?'

There will probably be some correlation, but we'd expect that it was the location where the officer was then they conducted the search. Please let us know in your report if this is not what the data suggests.

Regarding the analysis of whether there is improper behavior regarding clothing to be removed, should all the protected classes be analysed?
We've had issues in the press suggesting these are more prevalent with women of some ages, so please check across gender, age and ethnicity.

There seem to be stations which aren't reporting clothes removal information at all. How should we deal with this?
Thank you for bringing this to our attention. Please detail this in your report so that we can take any necessary action.

Is there any reason for some days to have much more stop & search than the majority of days? Any special events or specific operations?
This may be related to specific policing operations, but we do not have any insider information on this. Please mention it in your report.

What is the meaning behind legislation? Could you provide some context for this?
The legislation field is the legislation under which the search was performed. It should in theory never be missing, but from what I understand it has been found to be missing in the dataset. Please let us know about this in your report.

2. Data analysis

2.1. General analysis

The dataset is composed of 660610 observations (stop and search operations).
After an initial clean up it was reduced to 308129 observations .

Each observation have the following information about the search of operation:

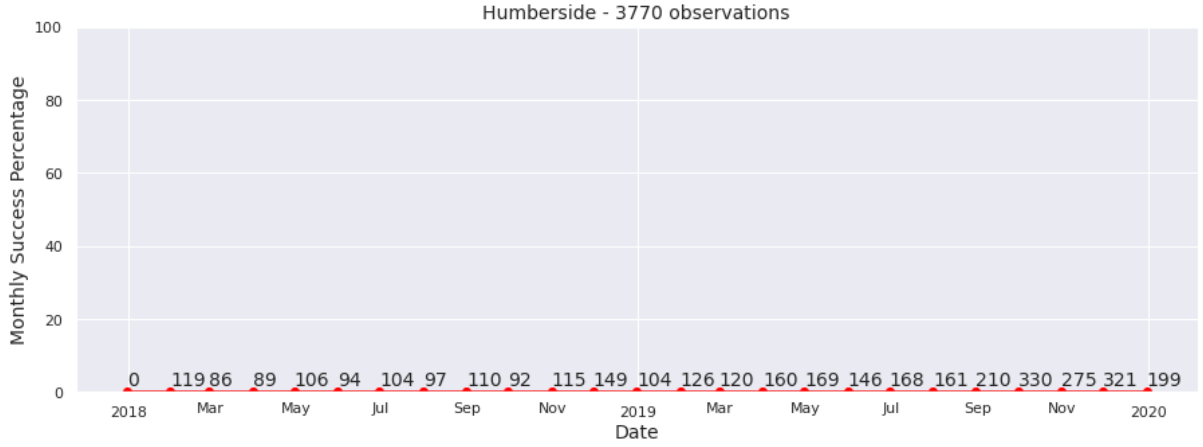
- Date
- Location
- Police station responsible for it
- Gender, Ethnicity and Age about the individual
- If the individual was asked to remove more than the outer clothing
- Legislation used to justify the operation and the Objective of that search
- Outcome of the search and if it was related to the Objective of the search

For details about each observation and dataset clean up see Dataset technical analysis in the Annexes.

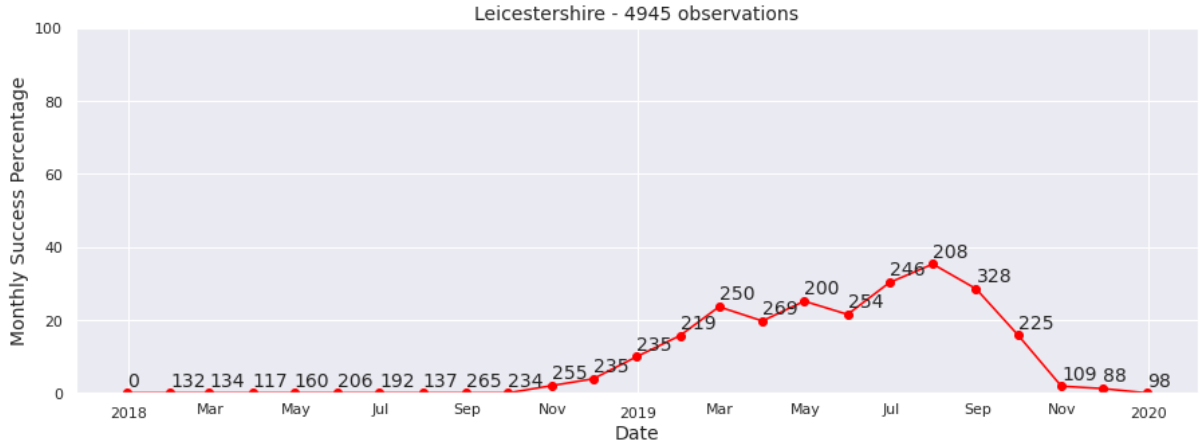
2.1.1. Search Rate in each station over time

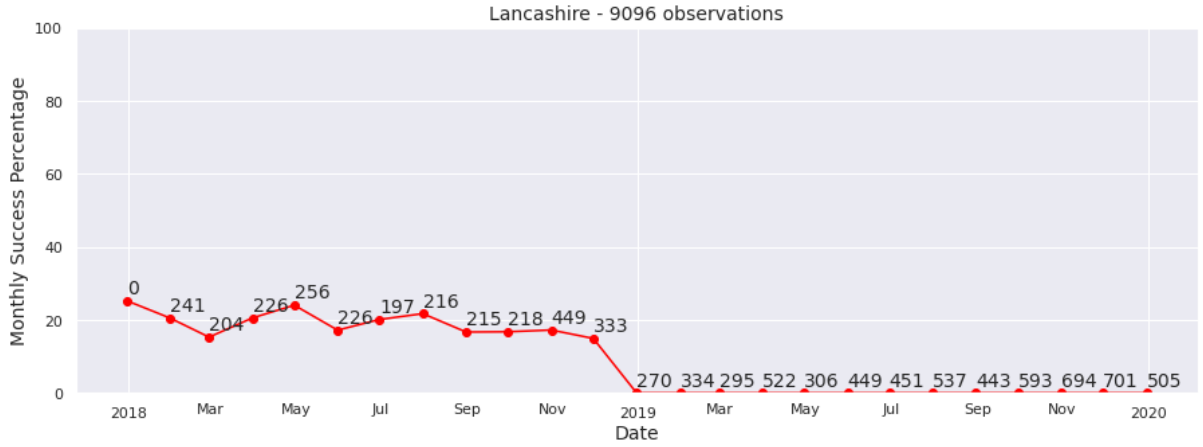
Missing data: Cambridgeshire have even more periods with missing data.

Zero success or data entry problems: Dyfed Powys and Gwent are in the same situation.

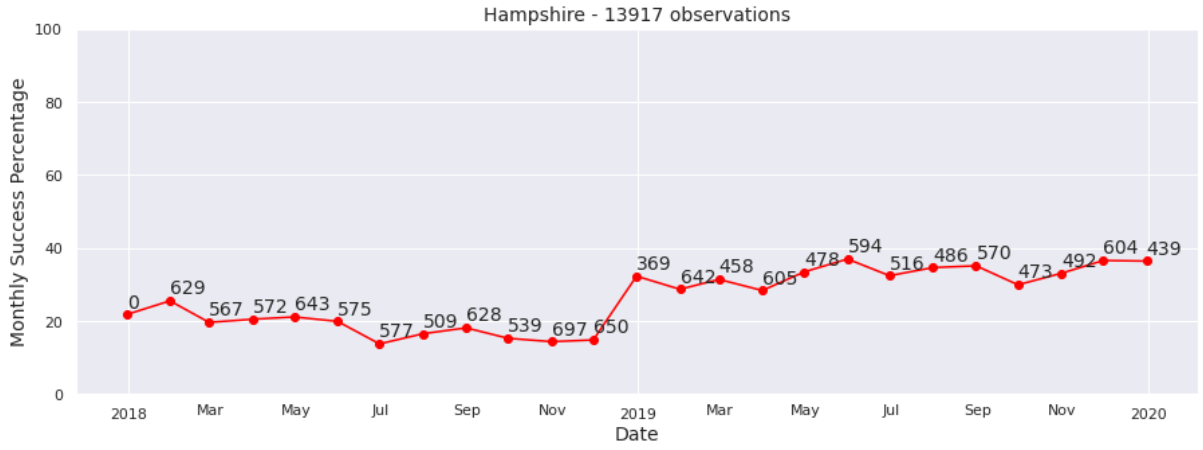


Continuous degradation: Leicestershire and Lancashire had some periods of success and then zero.



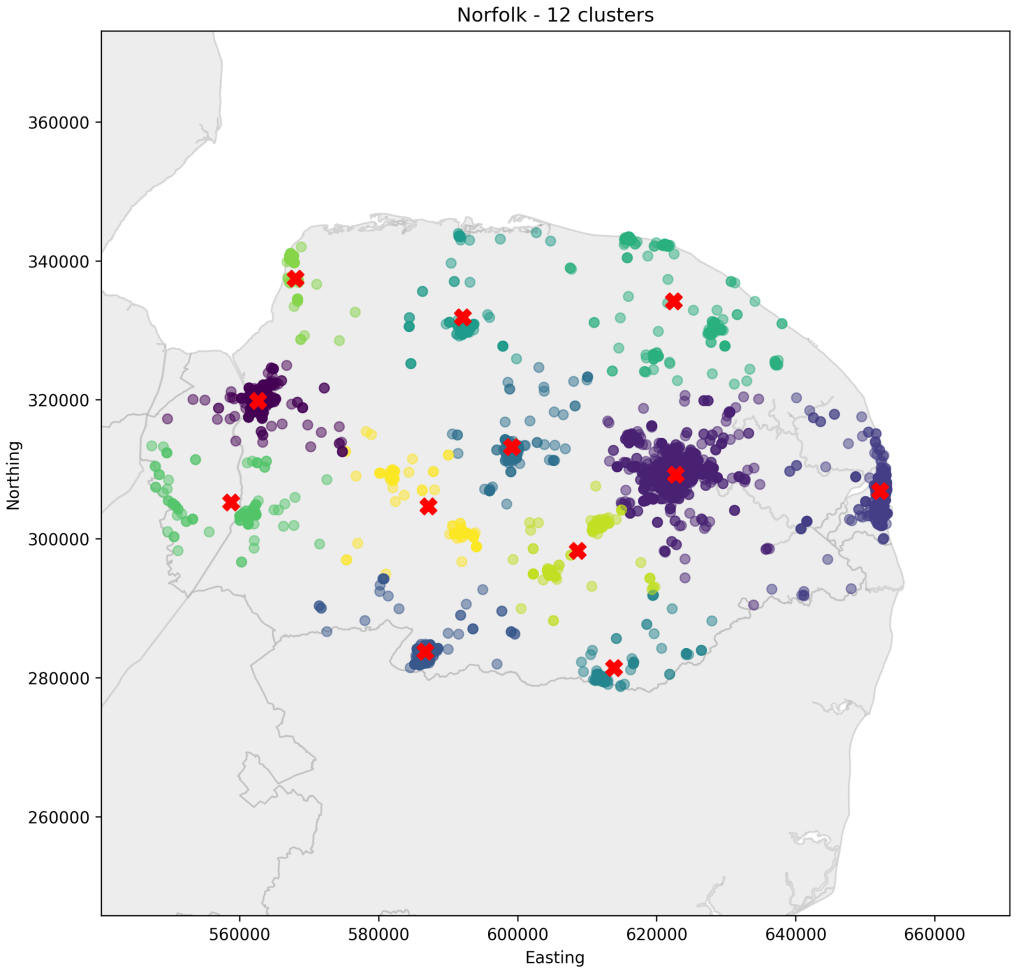


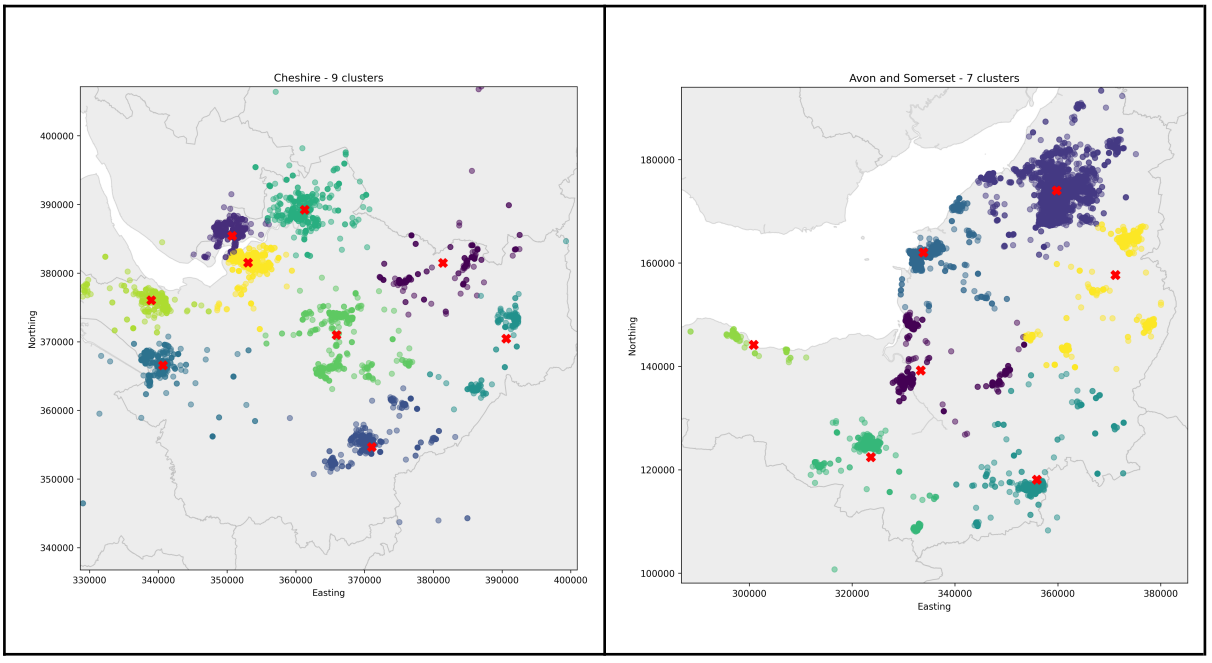
Continuous improvement: Hampshire is one of the stations improving over time.



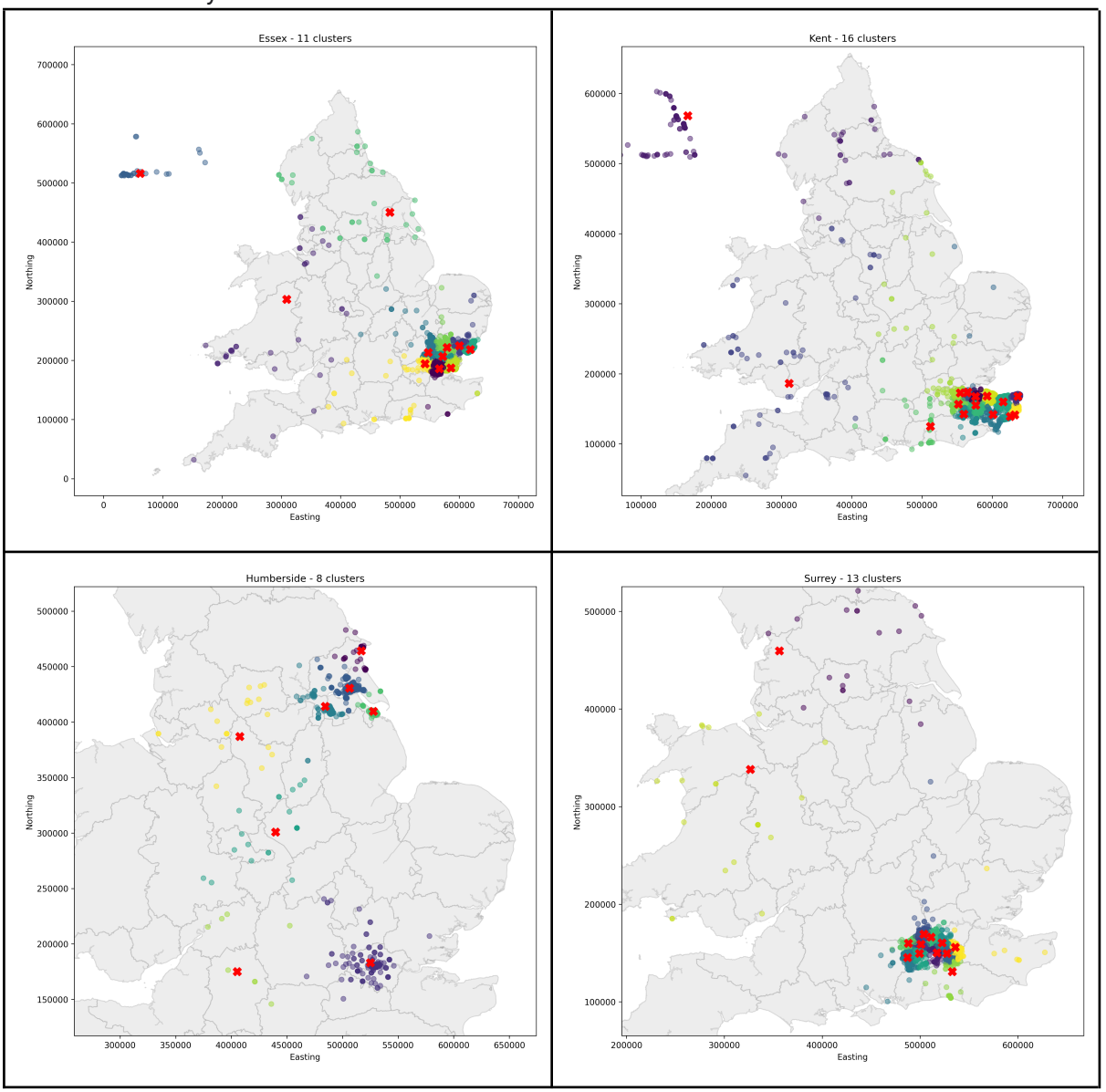
2.1.2. Station/Location comparison

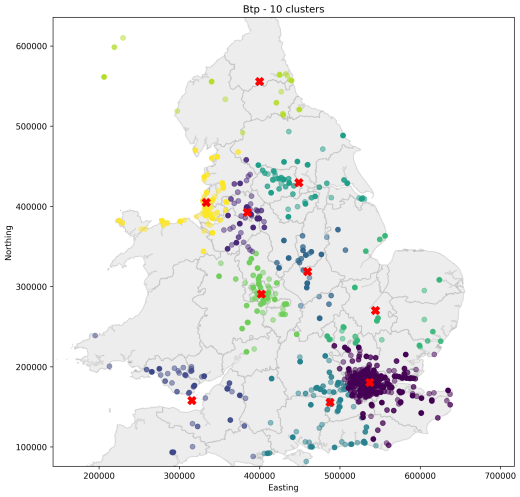
In general the location of the stop and search operations reside within the Police Station County area, with some occasionally done outside it.
Examples: Avon and Somerset, Cheshire, Norfolk.





Exceptions being: Surrey, Kent, Humberside, Essex, Btp, where some locations are more spread all over the country.

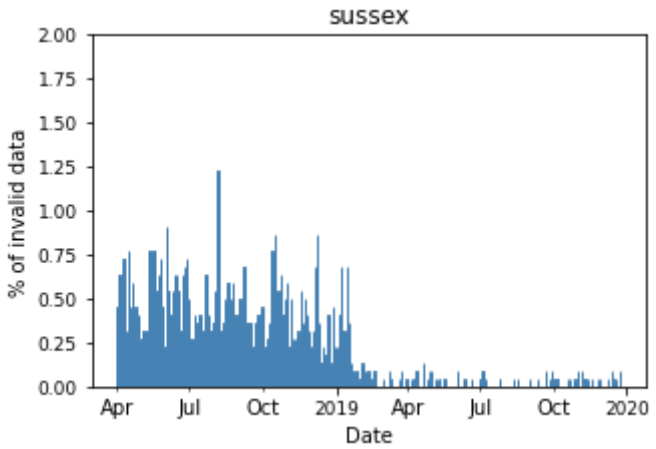




Btp was expected since it is the British Transport Police which polices railways and light-rail systems.

2.1.3. Data entry problems

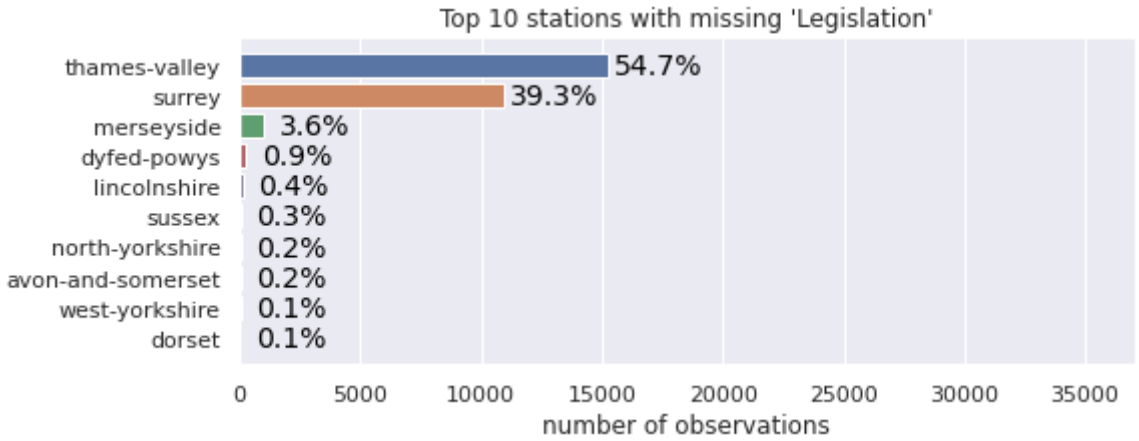
There are data entry problems in the observations with negative outcomes but considered as being related to the objective of search.
Doesn't seem to exist any specific pattern related to it, it is distributed throughout all stations over time and in general it is below 0.5% of the entry data.



Station Sussex is an interesting case where initially it was one of the highest percentage, and then some change happened in the beginning of 2019 that led to a significant decrease in wrong entries and becoming one with the lowest.

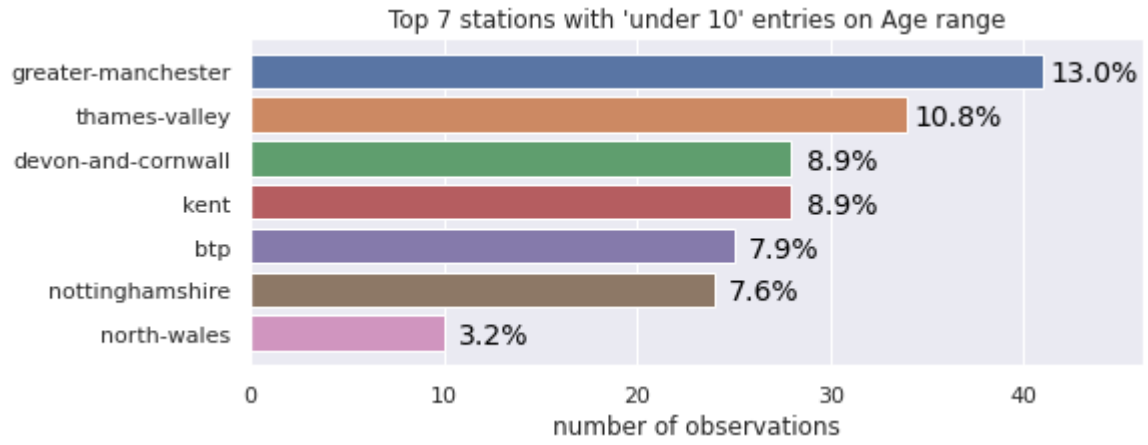
2.1.4. Missing Legislation data

There are 27908 missing entries about the Legislation used in the search operation.
The problem comes mainly from the stations Thames Valley and Surrey.



2.1.5. Subjects with ages unders 10

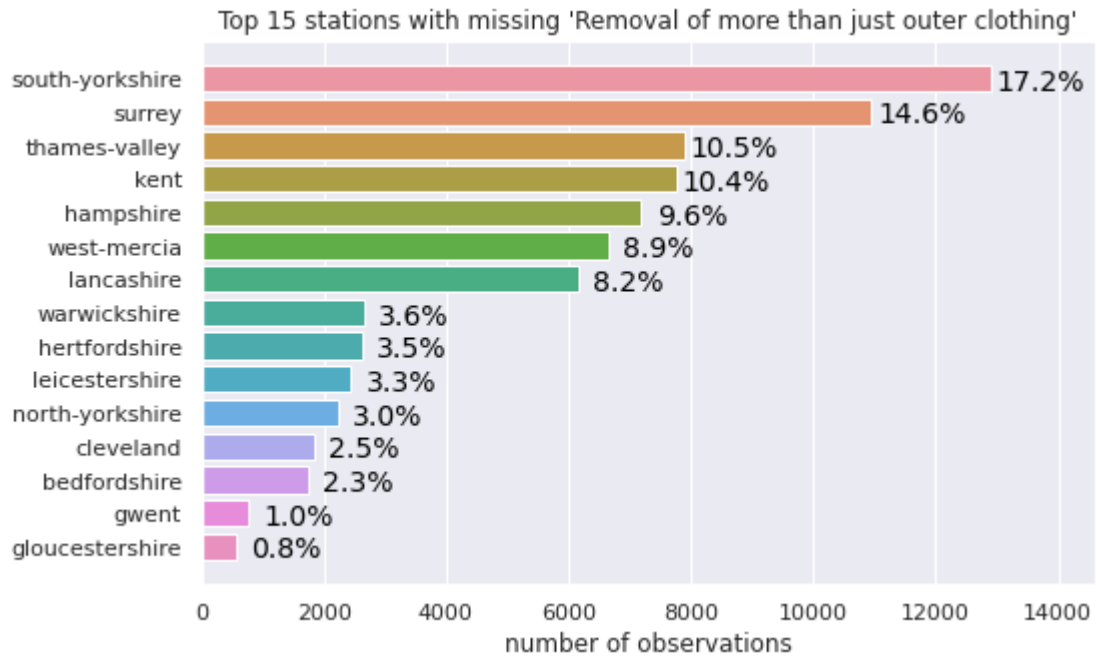
We found 316 data points where the identified ‘Age range’ of individual searched was ‘under 10’. This are the top stations that performed those entries:



Beside the high rate of entries on this set of stations, there doesn't seem to exist any more special pattern related to it.
It seems to be a wrong entry of the specific ‘Age range’ value and not a problem with the overall observation, given that the rest of the data in the observation seems to follow the general pattern in its distribution in relation to the rest of the data points in the dataset.

2.1.6. Missing clothes removal information across stations

There are 74991 observations with missing information about the ‘Removal of more than just outer clothing’.
This is the distribution of the missing data across stations:



2.2. Business questions analysis

Our analysis had the objective to quantify the degree of:

- discrimination in search success rate
- discrepancy in asking for the removal of more than just clothing

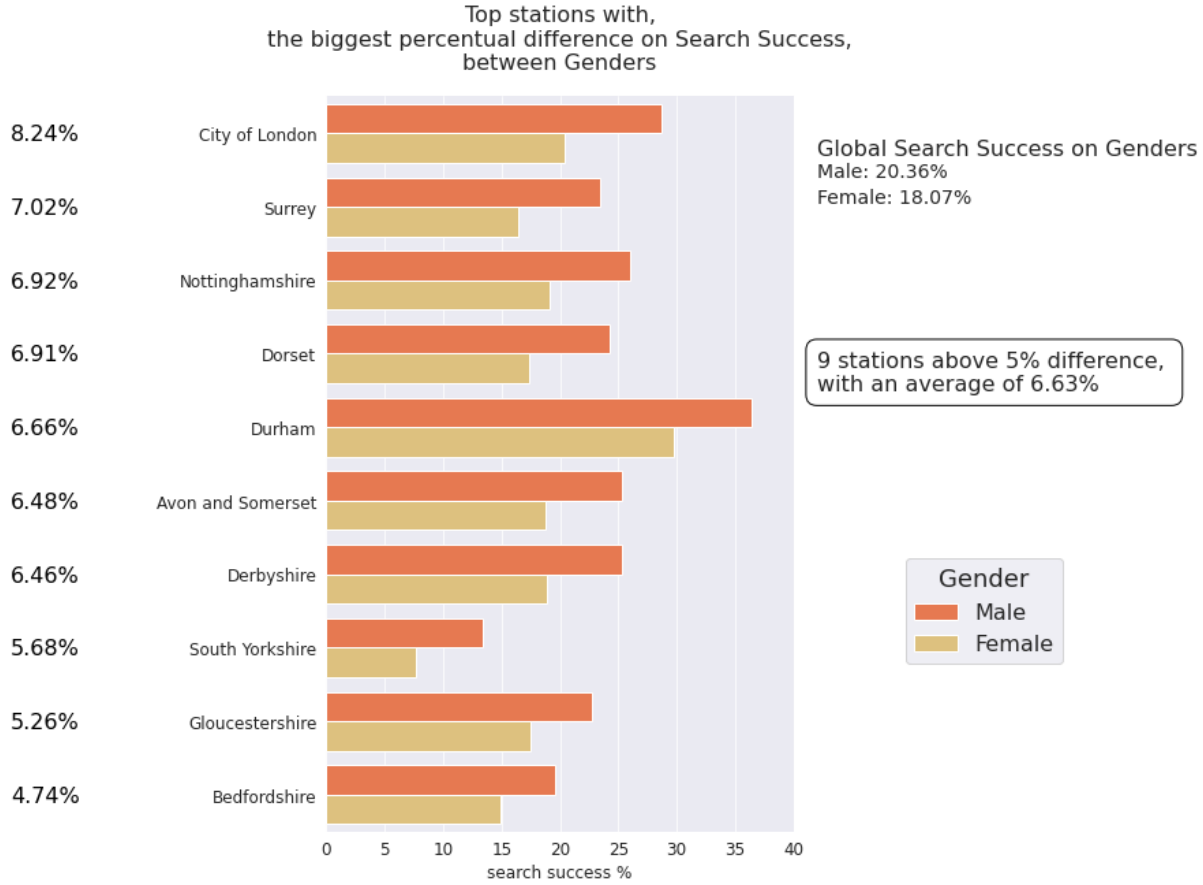
between Gender, Ethnicity and Age on each station and globally.

2.2.1. Discrimination in search success rate

We will analyse the existing degree of discrimination by comparing the search success rate between Genders and between Ethnicities on each station.

2.2.1.1. Gender

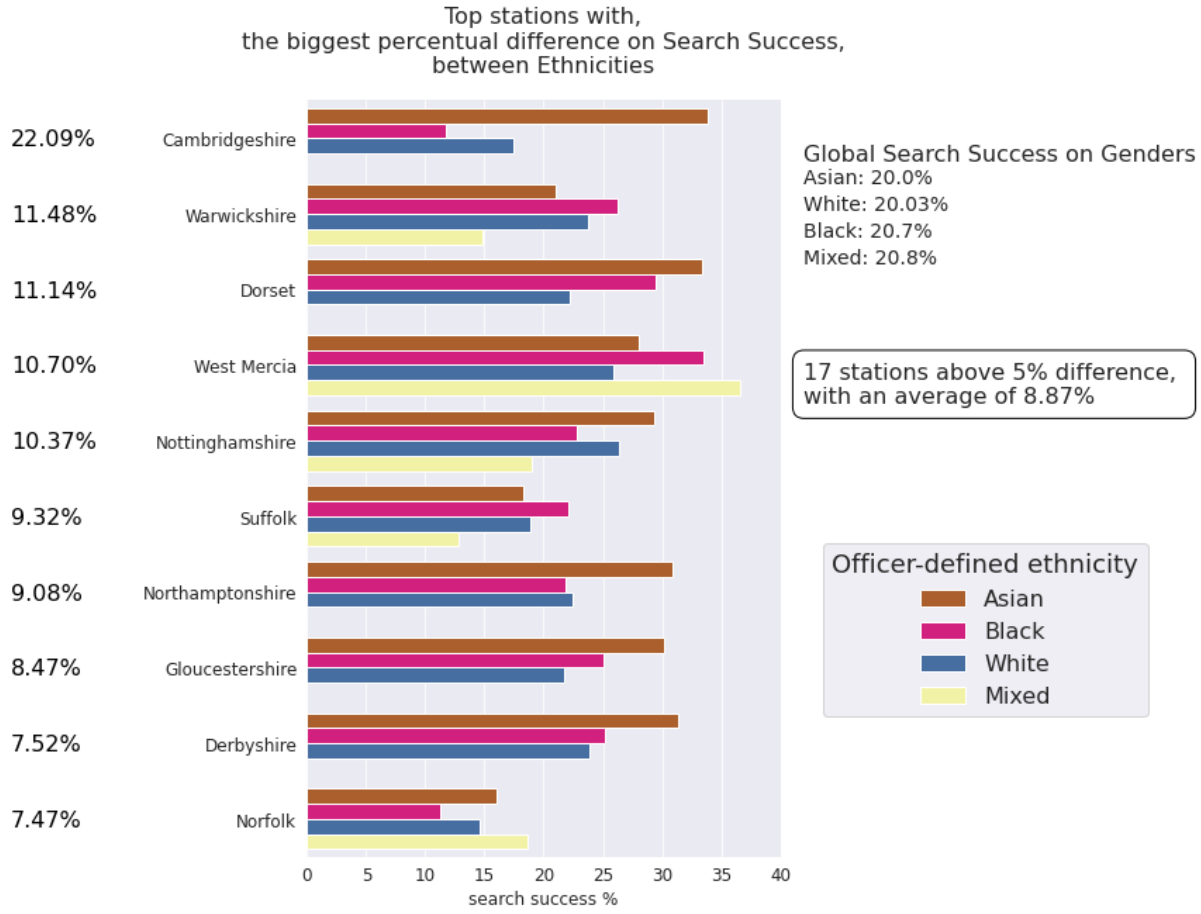
There seems to exist a general tendency for search operations on Male individuals to have higher success rates (2.29% on the global average). ~~success percentage~~
This is the top 10 stations where that difference is more accentuated:



2.2.1.2.

2.2.1.3. Ethnicity

Globally the search success rate between all Ethnicities seems to be well balanced - the difference is below 1%. But looking at the station level, the top cases of the biggest difference seem to show a tendency for individuals of the Asian Ethnicity to have higher search success rate than the others.

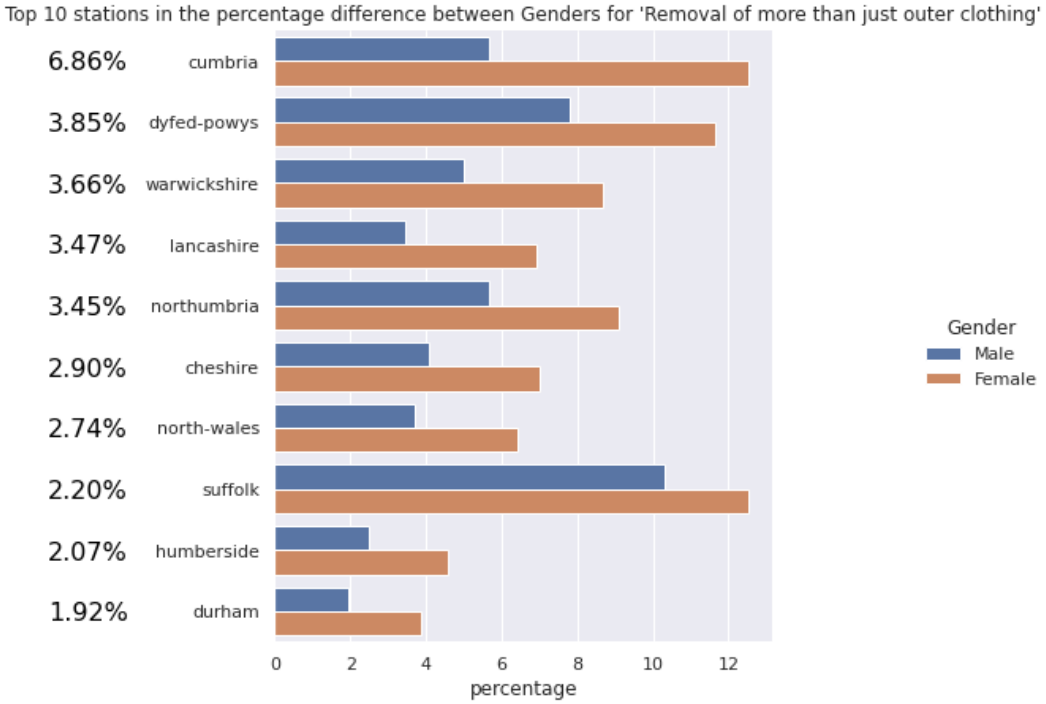


2.2.2. Discrepancies on cloth removal

We will investigate the discrepancies on asking for the removal of more than just the outer clothing by comparing its rates between Genders and between Ethnicities on each station. In the end we will compare globally each Gender in (Ethnicity, Age) groups to see if any sub-group is especially targeted in relation to the others.

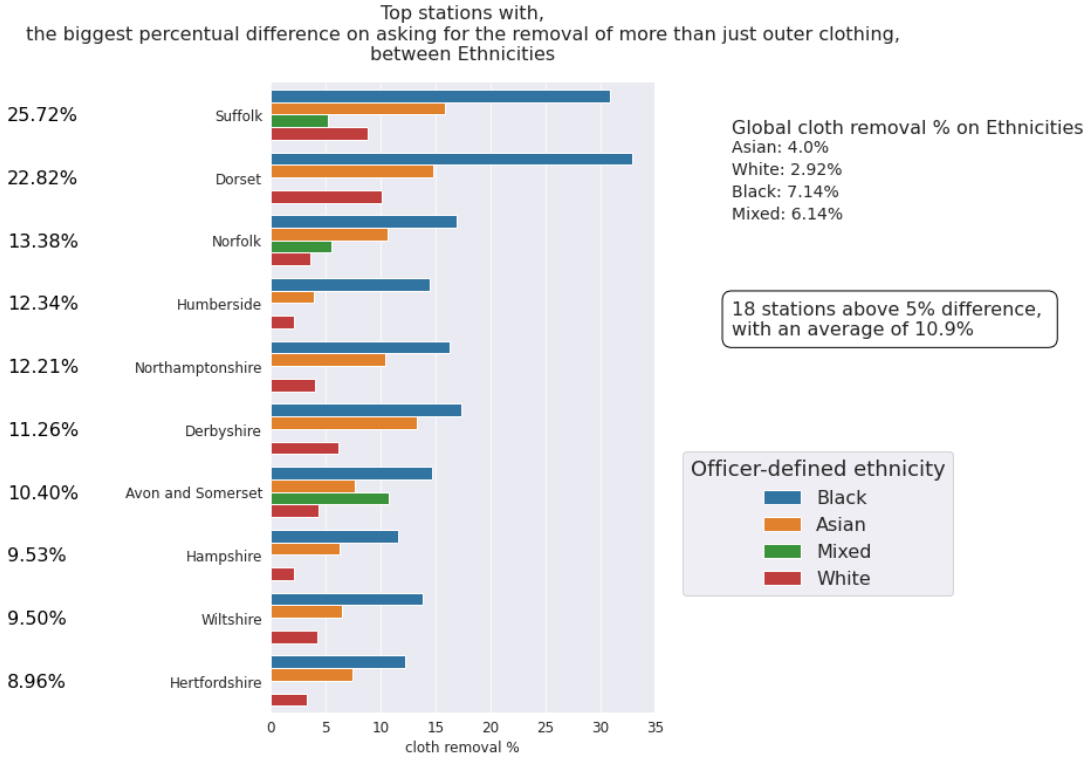
2.2.2.1. Gender

Data shows how individuals from the Female Gender are systematically more frequently asked for cloth removal, especially in Cumbria station.



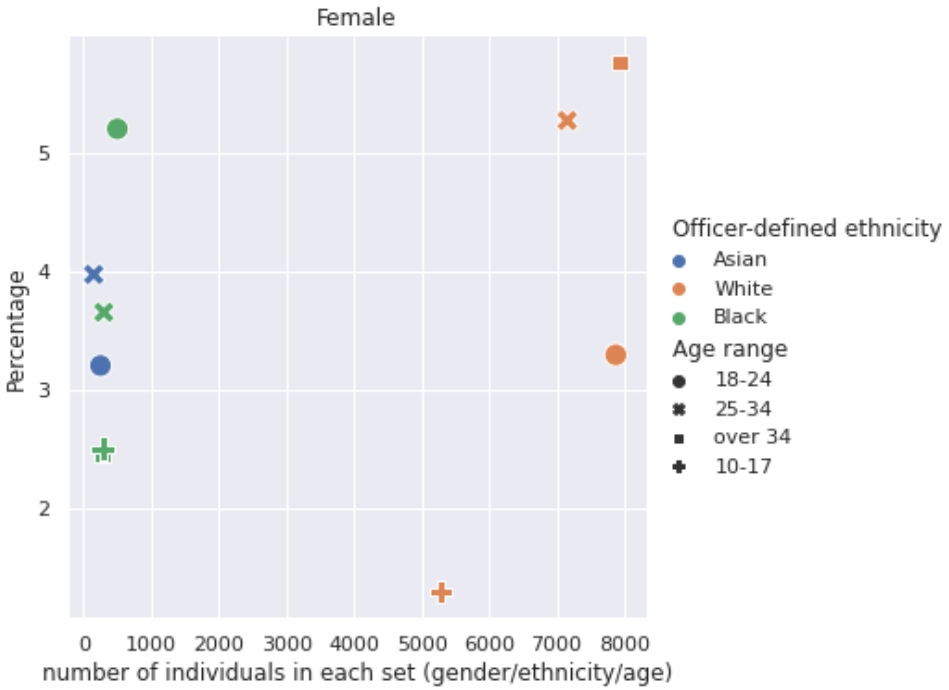
2.2.2.2. Ethnicity

Globally Black and Mixed ethnicities are 2 times more frequently asked to remove more than the outer clothing. Looking at each station individually, we can see how in general individuals of White ethnicity are asked for the removal of more than the outer clothing less than 5% of the operations, while individuals of Black ethnicity range from 12 to 30%.

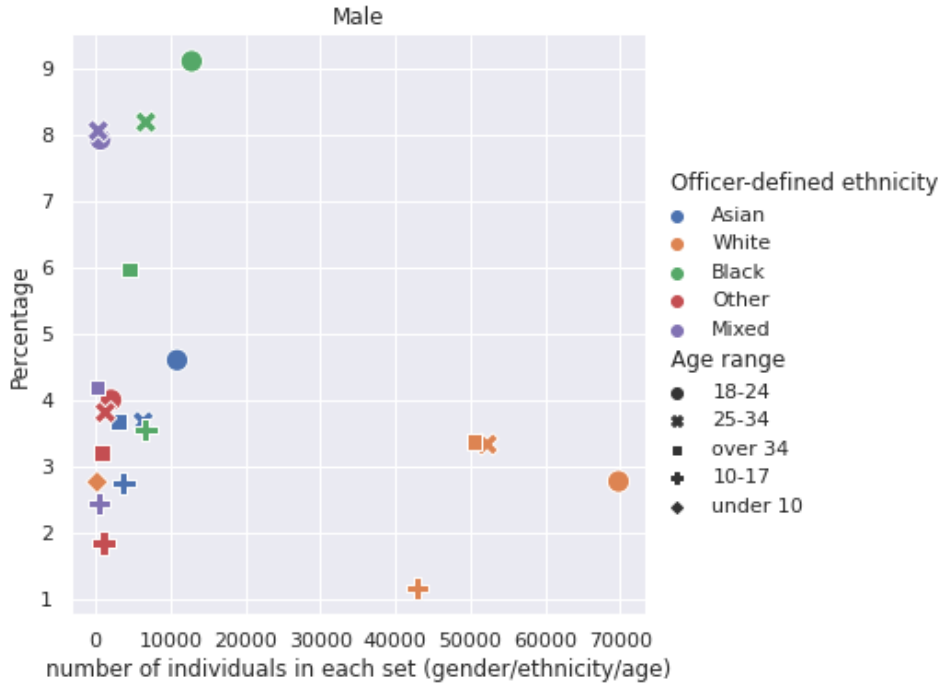


2.2.2.3. Gender/Ethnicity/Age sets globally

Individuals belonging to the sub-group (White, Female, age over 34) have a 2.18% higher rate of being searched than the median (4.03%).
The sub-group (Black, Female, age 18-24) have 1.62% higher rate but it's based on only 500 observations.



Individuals identified as Black and Mixed Males in the age range 18-24 and 25-34 are being asked for cloth removal 2 times more than the median across Male sub-populations (4.3%). In the case of Mixed Ethnicity there is significantly less data. More data in the future will be useful in clarifying the tendency.



2.3. Conclusions and Recommendations

We investigated the monthly search success rate on each station and saw cases of continuously improved(Hampshire) and continuous degradation (Leicestershire, Lancashire). It also came to our attention cases of missing data (Merseyside, Cambridgeshire) and zero success rate on the searches (Dyfed Powys, Gwent, Humberside).

Then we identified cases of entry problems on the Age range and Legislation on various stations, and some mismatches between the searches location and station location (Surrey, Kent, Humberside, Essex) which might be worth understanding.

Regarding the issue of discrimination, Male Gender and Asian Ethinicitiy (Cambridgeshire) seem to have tendency for higher search success rates when compared to the others.

In the case of cloth removal, the Female Gender (Cumbria) are Black Ethnicity (Suffolk, Dorset) are systematically more targeted. When analysing between Gender/Ethnicity/Age sub-groups globally, individuals identified as Black and Mixed Males around the age of 18-24 and 25-34 had generally 2 times more frequency than the average across Male sub-populations.

3. Modeling

3.1. Model expected outcomes overview

Based on the available data and model performance, it's expected that of all the received observations, only 45% will be given green light, reducing in more than 50% the search operations. This reduction comes at the cost of missing around 6% of successful searches. Of all green light operations, it is expected that 29% will be successful, an improvement of 8% in relation to the current success.

Evaluating the success per station and per search objective, the 10% requirement is expected to be met, with the possible exception of 'Article for use in theft' on the station Cleveland (4%) and Northumbria (9%). By considering a minimum of 100 observations, Cleveland wouldn't be accounted for.

Regarding the maximum variation of 10% between station's success rate, of the 38 stations considered, 29 should fall between 22% and 32% success rate. From the 9 stations outside, 3 are below and 6 above. The complete list is presented on the Annexes.
Even though this requirement isn't being completely met, there is margin for improvement with better model tuning.

Balacing the degree of discrimination between Genders and Ethnicities seems to be the area with the worse performace. In relation Genders, we expect 24 stations with discrimination above 5% with an average of 12%. Between Ethnicities it's expected 25 stations with discrimination above 5% with also an average of 12%.
These averages are being affected by extreme outliers, so an increase on the insignificant data threshold from 30 might help.
Besides that, it would be helpful to remove and investigate extreme outline stations, where for some reason there is 0% or 100% success on some class, resulting in a big discrepancy with any of the other classes. Examples are shown on the Annexes.
There is also room for improvement by better model tuning and trading-off with other requirements.

The model's outcome will also depend on the characteristics of the new received data. If the new data has some inherent bias or has new values that didn't happen before, the model is expected to deviate from what's expected and new model iterations might be needed to adapt and correct it.

3.2. Model specifications

3.2.1. Clean dataset

See Dataset cleanup in the Annexes.

3.2.2. Feature engineering

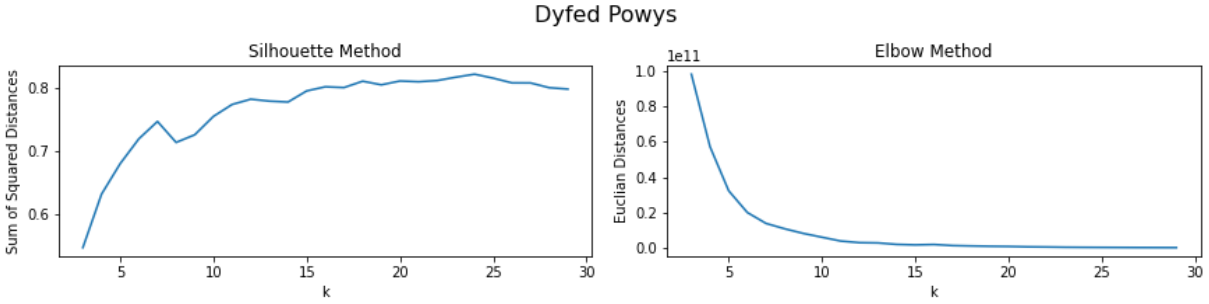
3.2.2.1. Date

Starting with the numeral features, from the column Date of each observation we will extract the (1) hour , (2) month and (3) day of the week. Those will be our 3 features and Date won't be used.

3.2.2.2. Latitude/Longitude

To make use of Longitude and Latitude, we created a point coordinate based on the combination of the 2, and then used the KMeans algorithm to classify the points from each station into various clusters. See K-means to find implementation in various languages.

To find the near-optimal number of clusters of each station, we used a combination of Silhouette Method and Elbow Method.
Example on Dyfed Powys station:



This is manual work for each station after printing the 2 plots for each station. In the Silhouette Method, we want the highest Sum of Squared Distances, wherein the Elbow method we want the lowest Euclidean Distance. In this case of the station Dyfed Powys, k=17 is a very optimal number. n=13 is probably already a good enough k.
See optimal_pngs folder for the chosen k on each station.

After having a good enough k on each station, we can train the KMeans on the points, and it will predict a cluster number for each Longitude/Latitude point pair. Every station will have a cluster number from 0 to k, but the number itself is just for identification of the cluster and doesn't have any intrinsic meaning, so there is no interest in using it as a numeric feature. We also want to differentiate the cluster 3 of a station to the cluster 3 of another station, so we will attach the name of the station to the cluster number, ex: dyfed-powys3, sussex3,... After all points being classified, this column is ready to be used as a categorical feature - (4) cluster.

3.2.2.3.

3.2.2.4. Categorical features

The other categorical features will be: (5) Object of search , (6) Legislation , (7) Type , (8) Part of a policing operation, (9) station.
'Gender' and 'Officer-Defined Ethnicity' were not used with the intention of reducing the discrimination at the cost of a little improvement in the predictions.
'Removal of more than just outer clothing' and 'Age range' won't be used to decide wheter or not to do a search, only for discrimination analysis.

3.2.2.5. Category encoding

To encode the categorical features we will use OneHotEncoding, which will create a new feature for each unique value on the column, and assign the value 1 when it exists in the observation, otherwise 0. This will ensure that the model won't give inherent meaning to each Category numberations. We don't want ranking between the values of any features.

In the case of the clusters there are more than 400 uniques. To avoid creating 400 new features we used LeaveOneOutEncoding where each value in the column is replaced with the mean target value for that category, but excluding the current observation value. This way each category value will have a range of close values on not the single mean number, improving the generalization when compared to TargetEncoding.
For more detailed information and how to implement it in other languages see: OneHotEncoding and LeaveOneOutEncondig.

3.2.2.6. Filling missing values

There will be null values on the 'cluster' and 'Legislation' features, and we will fill them with a constant.

3.2.3. Classifier and scoring

The used classifier is called XGBoost. It is a gradient boosted tree ensemble algorithm and can be used for regression or classification problems. Visit this to see how to implement it in various programming languages.
We tried different learning_rate, max_deepth, alpha and lambda regularizations, but at the time it didn't seem to improve the predictions so in the end the default configurations were used. With a more clear account about the costs of each metric, some tuning like regularization might be beneficial and lead to good trade-offs.See this for more information about XGBoost tuning.

After making the predictions using the training and target data, the model was evaluated with the AUC Score which uses the predicted probabilities by the classifier. It will be around 0.68, and doing cross_validation with 5 folds gives an average of 0.67 AUC score, suggesting that it is stable and generalizing well.

Finally to select the ideal threshold when transforming the predicted probability into a classification of go/no-go to the search operation, we set the value to 0.228. See Annexes for more details on this tuning.

3.3. Analysis of expected outcomes based on training set

- 29% precision
- 65% recall
- 12% true positives
- 12% global discrimination

3.4. Alternatives considered

- Different types of encoding variables
- Different feature selection to reduce discrimination
- Different feature selection to increase performance
- Different classifiers
- Different classifier tuning
- Different number of clusters in each station
- Different thresholds for successful classifications based on predicted probability

Model	Baseline	Second iteration	Third iteration	Best model
Requirement 1 - success rate	0.33	0.27	0.29	1
Requirement 2 - global discrimination (gender)	0.22	0.12	0.12	2=3
Requirement 2 - global discrimination (ethnicity)	0.18	0.10	0.12	2
Requirement 2 - nº of station discrimination above 5% (sex)	26	24	24	2=3
Requirement 2 - nº of station discrimination above 5% (ethnicity)	30	22	25	2

3.5. Known issues and risks

- Model may not respond well to unseen stations since the station features are very important to the model performance.
- Data normalization is missing. Being case sensitive, new entries that differ on the known values by a simple upper case will be considered a new category.
- A way to convert Latitude and Longitude data into the best cluster isn't implemented yet. As a result, all new data won't have cluster information and take advantage of Latitude and Longitude, while the model was trained with that information. Based on this it will probably underperform.
- The model is lacking good strategies to impute missing data, particularly on Latitude/Longitude and Legislation.

In general there isn't good enough processing to be able to readily categorize new unseen data and use it in the model. If new observations fail during some prediction, it will be saved on the database so that it can be analysed and the problem fixed on next iterations.

4. Model Deployment

4.1. Deployment specifications

The model is integrated in a server using Flask, which is deployed with Docker on Heroku. This are the steps to replicate the model deployment:

1) create a repository on github with the project (includes server file app.py, docker file docker (Dockerfile) and heroku file (heroku.yml).
2) create an app on Heroku and a Postgres database
3) install Heroku CLI and login on the project folder
4) add a git remote to heroku: heroku git:remote -a name_of_the_heroku_app
5) run heroku stack:set container to be able to use python dependencies on heroku
6) push to heroku: git push heroku master

The server runs a flask application with a REST API with the 2 following endpoints:

1) **should_search/**
Receives a JSON request with a search observation with following fields:

```
{
 "observation_id": <string>,
 "Type": <string>,
 "Date": <string>,
 "Part of a policing operation": <boolean>,
 "Latitude": <float>,
 "Longitude": <float>,
 "Gender": <string>,
 "Age range": <string>,
 "Officer-defined ethnicity": <string>,
 "Legislation": <string>,
 "Object of search": <string>,
 "station": <string>
}
```

If the observation ID already exists on the database, it returns an error message.
Otherwise the application transforms the JSON into the appropriate format so that it can be fed to the model and generate a True or False prediction of whether or not the officer should do the search.

The information is then stored on a PostgreSQL database table.
The table contains a column for:
- entry ID (id - serial),
- entry Date (date_received - timestamp)
- observation id (observation_id - text)
- observation body in the JSON format (observation - text)

- predicted outcome (predicted_outcome - boolean)
- true outcome (outcome - boolean)

The true outcome column is always saved with a null value, it can only be filled through the next endpoint (search_result/).

Finally, if all goes well, a 200 response is returned with the predicted outcome in the format:
{
"outcome": <boolean>
}

2) **search_result/**
This endpoint allows us to input the true outcome of a search by receiving a JSON request with following fields:
{
  "observation_id": <string>,
  "outcome": <boolean>
}

If the “observation_id” doesn’t exist on the database an error message is displayed, otherwise the application updates the “true outcome” column of the received “observation_id” to the received “outcome”.
After that the application returns a code 200 response with the “observation_id”, the “true outcome” and the “predicted outcome” that was stored on the database:
{
  "observation_id": <string>,
  "outcome": <boolean>,
  "predicted_outcome": <boolean>
}

4.2. Known issues and risks

In relation to the application, those are some of the existing problems:

- If the Latitude or Longitude fields are strings, an exception is thrown when trying to convert the value to float.
- When the station field is missing or it's a new one, an exception is thrown because a KMeans model to predict the cluster for that station isn't found.
- If the Date is in a bad format that can be inferred (see infer_datetime_format), an exception is thrown.

Those 3 cases will lead to a 500 server error response and can be fixed in future app versions.
In the case of the JSON request being in the wrong format: ex: 'True' instead of 'true' or empty fields, the server will respond with a 400 bad request code.

In relation to Heroku, since we are using the free version, there are various kinds of limitations, including the application container going to sleep after some time idle, and the PostgreSQL database having a limit of 10.000 entries in the table.
The free version doesn't have threshold alerts so we will have to keep track of the current state in order to avoid the limits being reached.
For more information on the limitations see: https://www.heroku.com/pricing

5. Annexes

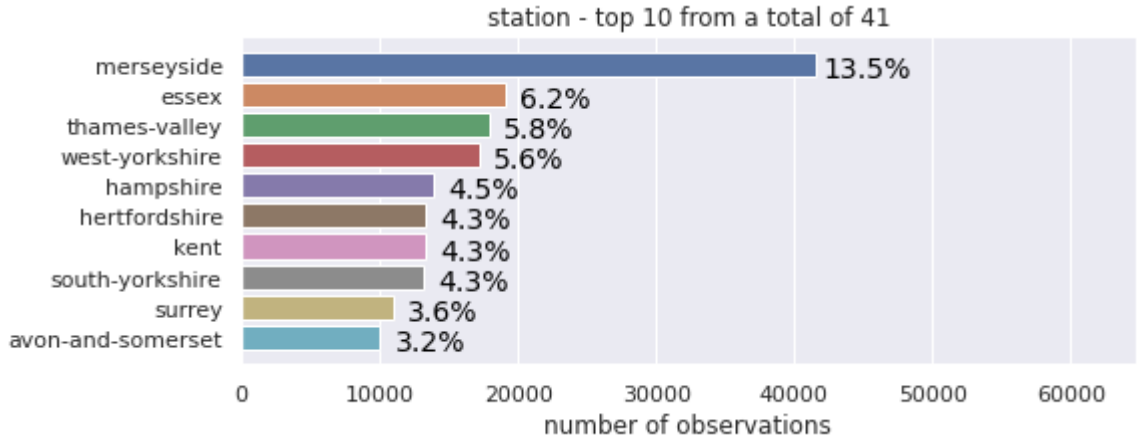
5.1. Dataset technical analysis

The initial dataset was composed of 660610 observations (stop and search operations). The dataset analysis was done after a clean up that reduced it to 308129 observations.

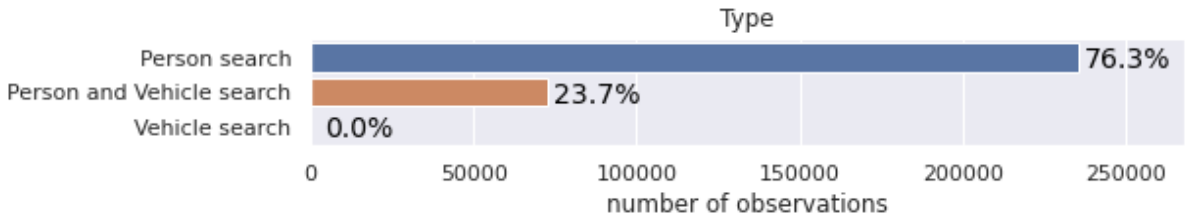
5.1.1. Dataset overview

Each observation is composed with the following information:

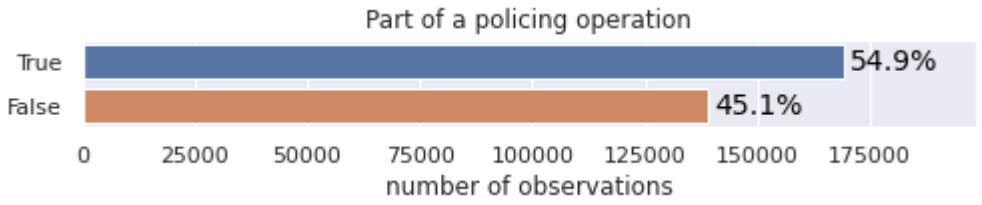
- **Observation_id**: search operation identifier
- **Date**: date of the observation. Searches cover the period from Dec 2017 to Dec 2019
- **Latitude**: coordinate of the search location, ex: 51.540219
- **Longitude**: coordinate of the search location, ex: -1.764708
- **Station**: police station where the search belongs to



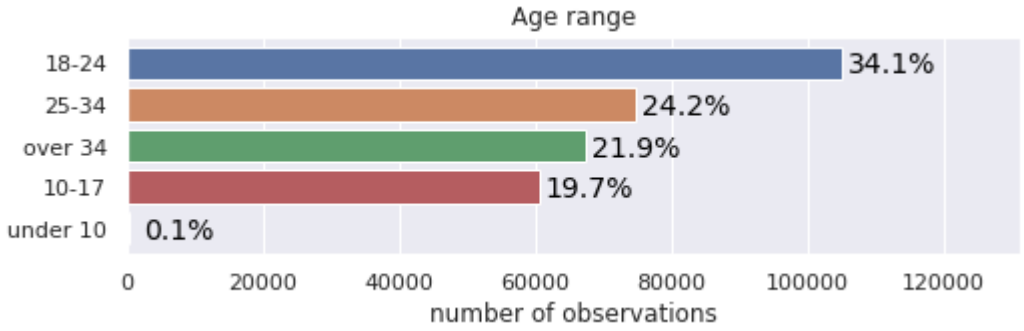
- **Type**: if it was a person, vehicle or person/vehicle search



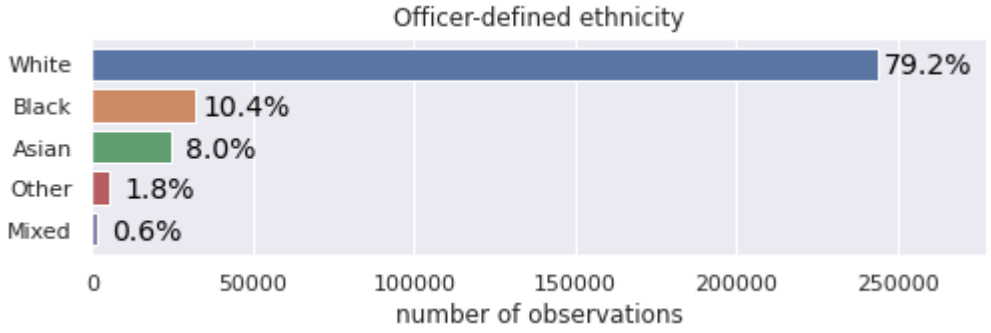
- **Part of a policing operation**: if the search was part of a policing operation



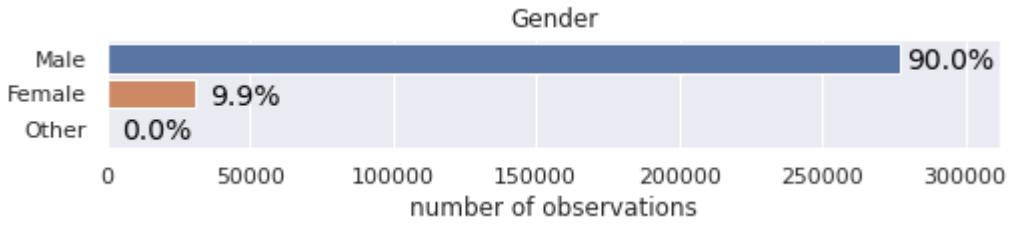
- **Age range**



- **Officer-defined ethnicity**: ethnicity of the individual searched defined by the officer

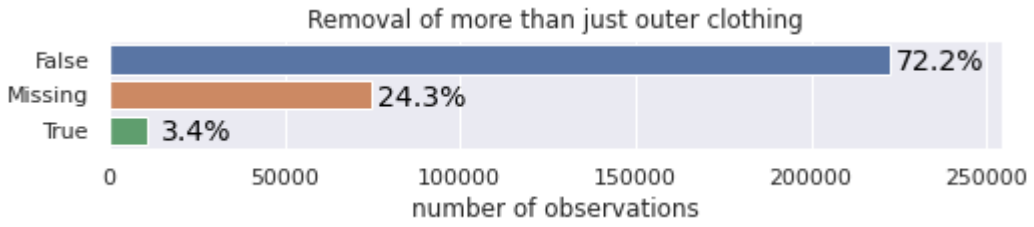


- **Gender**

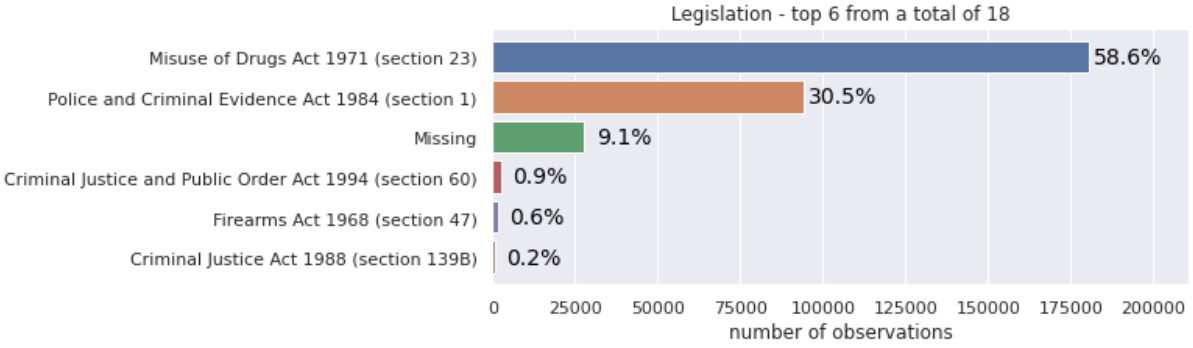


-

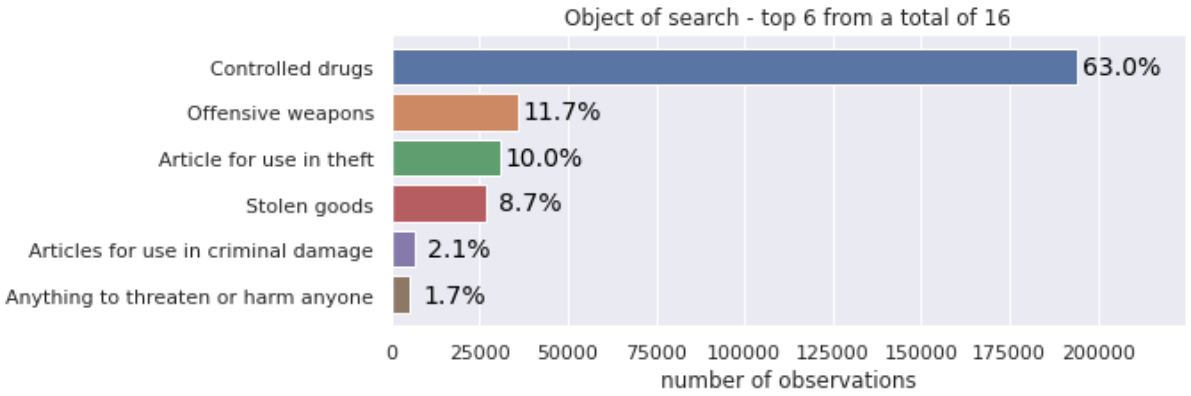
- **Removal of more than just outer clothing**



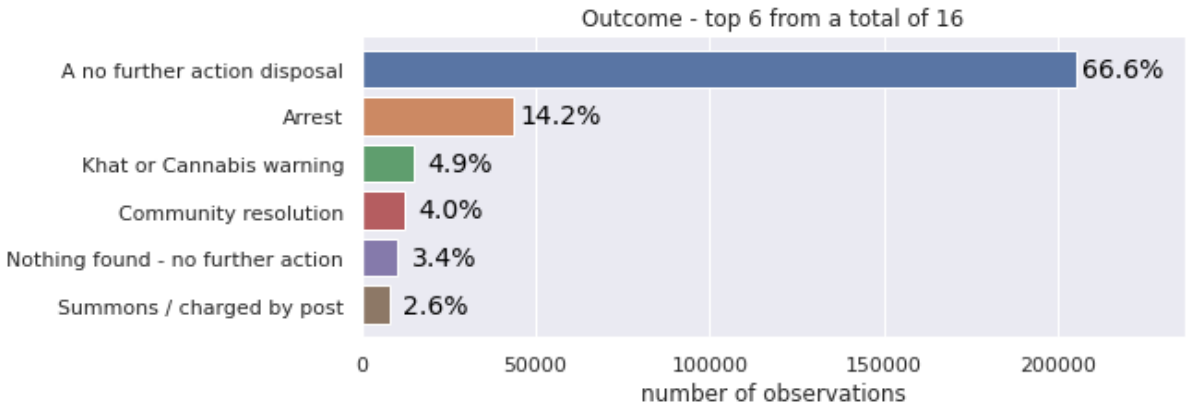
- **Legislation**: law used to justify the search operation



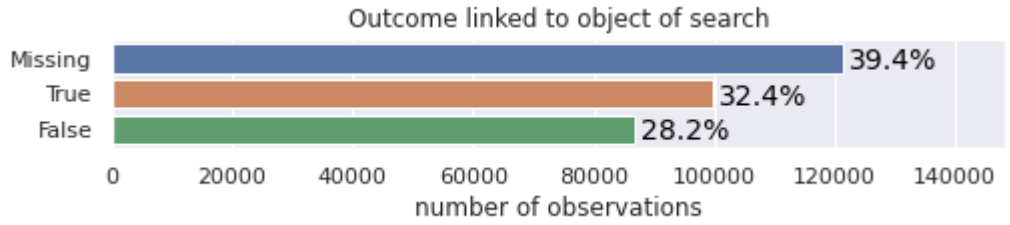
- **Object of search**: objective of the search operation



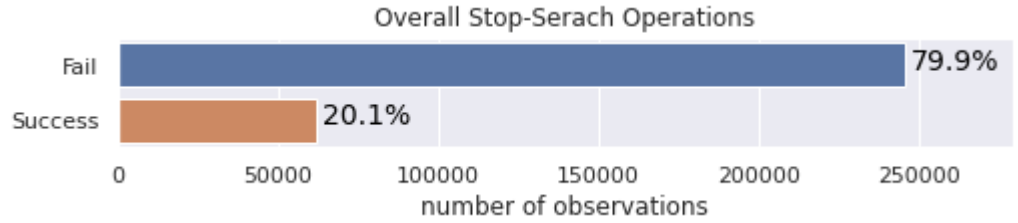
- **Outcome**: result of the search operation



- **Outcome linked to object of search**: if the result matches the expected



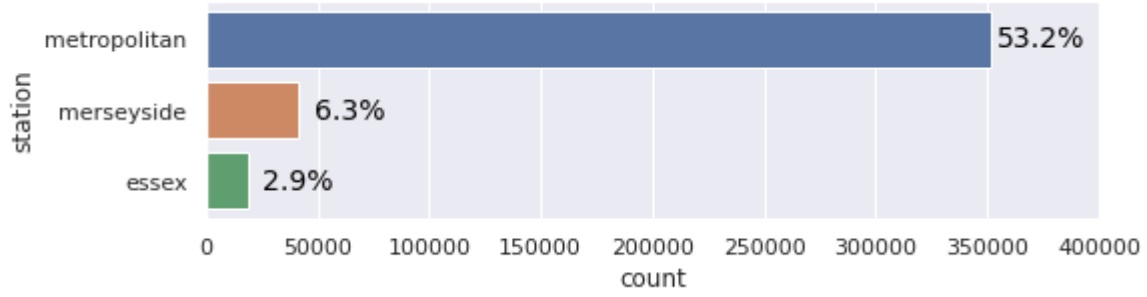
- An additional variable was created which combines **'Outcome'** and '**Outcome linked to object of search'** to define the success/fail of a stop and search operation:



5.1.2. Dataset cleanup

The dataset analysis was made after an initial data cleaning where the following operations were performed:

- All observations from the Metropolitan station were excluded because data about "Outcome linked to object of search" and "Removal of outer clothing" was always missing. Metropolitan was the station with the highest number of observations (351294 observations were removed).
  This was the initial dataset before the removal:



- The data combination of Station/Gender/Officer-defined ethnicity with less that 30 data points were excluded: 1188 observations.
- Missing data from 'Removal of more than just outer clothing' was considered False, except when it belonged to the Vehicle search type
- Missing data from Part of a policing operation was filled as False.

5.2. Business questions technical support

- Success rate on each station:

	station	precision
3	north-yorkshire	0.142857
4	north-wales	0.171875
5	wiltshire	0.200000
6	thames-valley	0.222222
7	cumbria	0.225564
8	merseyside	0.242748
9	norfolk	0.243243
10	suffolk	0.246094
11	lancashire	0.246377
12	west-yorkshire	0.251887
13	northumbria	0.262136
14	devon-and-cornwall	0.269966
15	dorset	0.271574
16	cleveland	0.272300
17	kent	0.276596
18	warwickshire	0.283820
19	lincolnshire	0.284047
20	btp	0.284545
21	greater-manchester	0.290576
22	avon-and-somerset	0.292022
23	sussex	0.292437
24	surrey	0.294682
25	gloucestershire	0.296029
26	staffordshire	0.299625
27	west-mercia	0.305808
28	nottinghamshire	0.306391
29	derbyshire	0.310526
30	hampshire	0.317372
31	northamptonshire	0.318966
32	hertfordshire	0.327575
33	bedfordshire	0.328829
34	essex	0.329630
35	cheshire	0.333333
36	city-of-london	0.340278
37	cambridgeshire	0.352941
38	leicestershire	0.362637
39	durham	0.373016
40	dyfed-powys	0.400000

- Examples of extreme variations in success rate between Gender/Ethnicities:

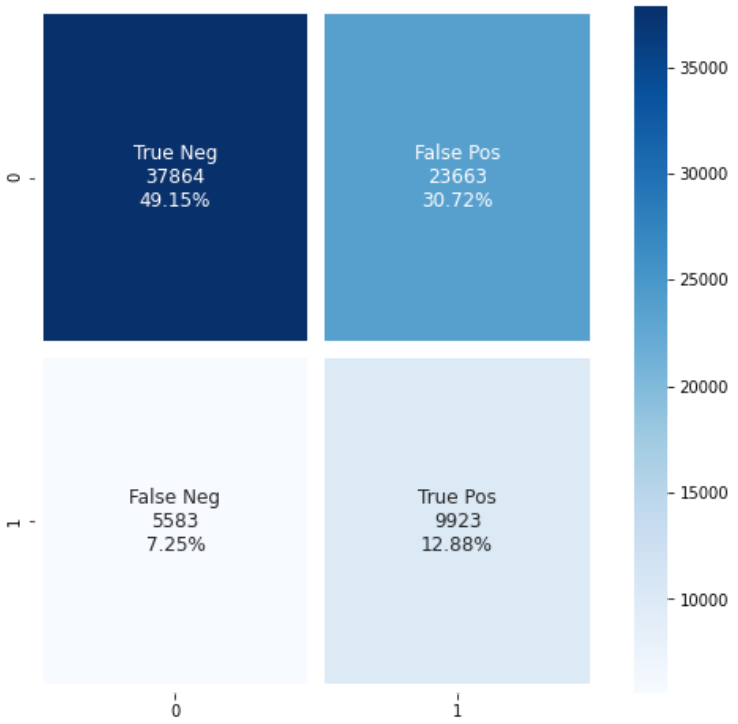
Lancashire
White: 25%, Black: 0%, Asian: 26%
Difference between Ethnicities: 26%

Dyfed Powys
Male: 33%, Female: 100%
Difference between Genders: 66%

5.3. Model technical analysis

Given the considered alternatives and resumed in the this table:

Model	Baseline	Second iteration	Third iteration	Best model
Requirement 1 - success rate	0.33	0.27	0.29	1
Requirement 2 - global discrimination (gender)	0.22	0.12	0.12	2=3
Requirement 2 - global discrimination (ethnicity)	0.18	0.10	0.12	2
Requirement 2 - n° of station discrimination above 5% (sex)	26	24	24	2=3
Requirement 2 - n° of station discrimination above 5% (ethnicity)	30	22	25	2



Third iteration was chosen because it is the most all-round model.
It performs almost as good as the Second in relation to discrimination, but is as a better overall precision.
It is also better tuned to reduce the False Positives with the intention that obvious searches that shouldn't be performed are excluded at the cost of missing some successful searches.