LISBON DATA SCIENCE ACADEMY

Stop and Search Operations in UK Policing

A Data Science Analysis and API Development for the United Kingdom Department of Police

Prepared for:
The United Kingdom Department of Police

Prepared by:
Francisco Oliveira
Data Scientist
Data Science Department

1 of April 2023

Table Of Contents

1. Client requirements

1.1 Summary

The United Kingdom Department of Police has requested that we investigate **accusations of discrimination** in their stop and search policy, **specifically targeting certain minorities and women of certain age groups and ethnicities**. Our task is to analyse the data and determine if there is evidence to support these claims. The client has also required that we create and host an API endpoint for authorising searches, which will be integrated into their police system for uniformity of decisions.

The analysis should **search for evidence of discrimination based on gender, ethnicity, and age** in terms of who is stopped and **who is asked to remove more than just outer clothing during searches**. We expect that some populations will exhibit higher levels of delinquency than others, but it shouldn't significantly affect the **success rate of searches between populations.**

After the analysis is completed, the client will run a proof of concept with our company for the use of the API in a few of their police stations. The main objective of the API is to ensure that searches are performed only when there is a **more than 10% likelihood of success**. It is also expected to **level the discovery rate without significantly diminishing the overall ability to detect offences**.

1.2 Requirements clarifications

The table 1.1 provides a clear delineation of the technical requirements derived from the original business requirements from section 1.1.

The success rate mentioned throughout the report is the number of searches that resulted in one of the positive outcomes listed and where the outcome was linked to the object of the search, divided by the total number of searches which in technical terms is the same as Precision score.

Table number 1.1

Business requirements	Technical requirements
Search for evidence of discrimination on gender and ethnicity across stations.	Precision score between different ethnicities/genders should not vary 5% in each station.
Verify if women of certain age groups and ethnicities are being asked more often to remove outer clothing.	Precision score when clothes are removed on women of different ethnicities/age-ranges should not vary 5% in each station.
Searches are performed only when there is a more than 10% likelihood of success.	Initiate a search only if the predicted probability of success is greater than or equal to 10%.
Level the discovery rate without significantly diminishing the overall ability to detect offences.	Maximise the recall rate to stay above 90% to detect the majority of offences, while ensuring that the level of discovery rate remains consistent across different population sub-groups. Specifically, aim to limit the discrepancy in precision scores between sub-groups to no more than 5% at each station and ensure that the discrepancy between stations, measured as the average precision score per station, does not exceed 10%.
Create and deploy a REST API.	Ensure the REST API remains operational and processes information as requested without any disruptions or errors.

The requirements related to success rates between different ethnicities/genders and when clothes are removed on women of different ethnicities/age-ranges not varying by more than 5% in each station may be challenging to achieve. It is possible that there are underlying factors that impact search success rates, such as crime rates in certain areas, which may make it difficult to control for all variables and ensure a consistent success rate across different populations.

Furthermore, it's important to note that the 10% threshold on the predicted probability might not necessarily be the optimal threshold for achieving the best performance. Different models may have different optimal thresholds, and it's possible that a better threshold could be found through experimentation and tuning. Therefore, finding the optimal threshold for the given model and dataset may require some trial and error and careful evaluation of performance metrics.

2. Dataset analysis

2.1 General analysis

The data provided contains information about **over 850,000 police searches conducted across England and Wales from 2019 to 2020**. This dataset provides a unique insight into the patterns and outcomes of police searches, which can be used to understand the **effectiveness of current policing practices and to identify areas for improvement**.

Understanding the patterns of police searches can help improve public trust and confidence in policing, particularly for marginalised communities that may be disproportionately affected by such actions by identifying **any potential biases or discriminatory practices in police search procedures**. This dataset, therefore, presents a valuable opportunity to explore these issues and work towards a more equitable and just society.

We will be focusing on the key features during this presentation, while more detailed information can be found in the annexes.

2.1.1 Type of search

The graph in figure 2.1 describes the type of search conducted by the police officers. The three possible types are Person search, Person and Vehicle search, and Vehicle search. There is a huge discrepancy between these values as “Person search” was by far the most common as seen below.

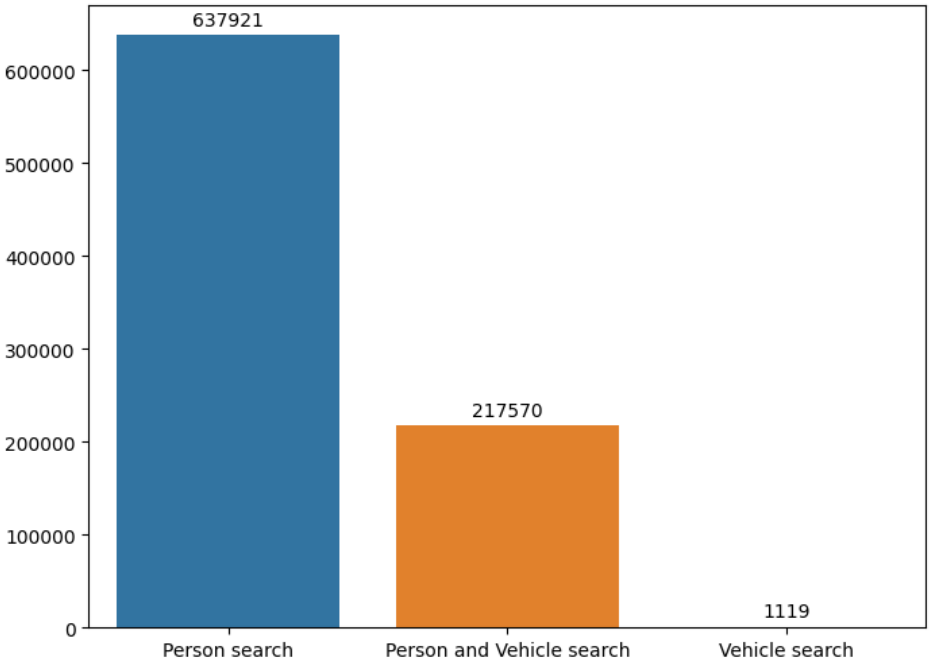

Figure number 2.1

2.1.2 Age

The data shows that a significant proportion of the searches were conducted on individuals between the ages of 18-24, with the second largest group being those aged 25-34 and then over 34. There were also a significant number of searches conducted on minors aged 10-17 and just a few under 10 as per figure 2.2.

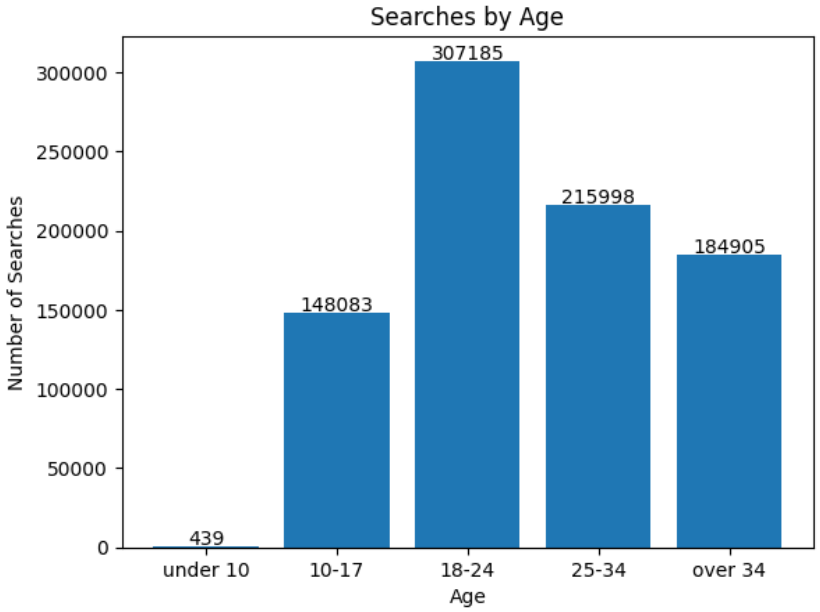

Figure 2.2

2.1.3 Gender

As seen in the figure 2.3 there is a big discrepancy between searches. The data suggests that gender may play a role in the likelihood of being searched, with males being more frequently searched than females or individuals who identify as another gender.

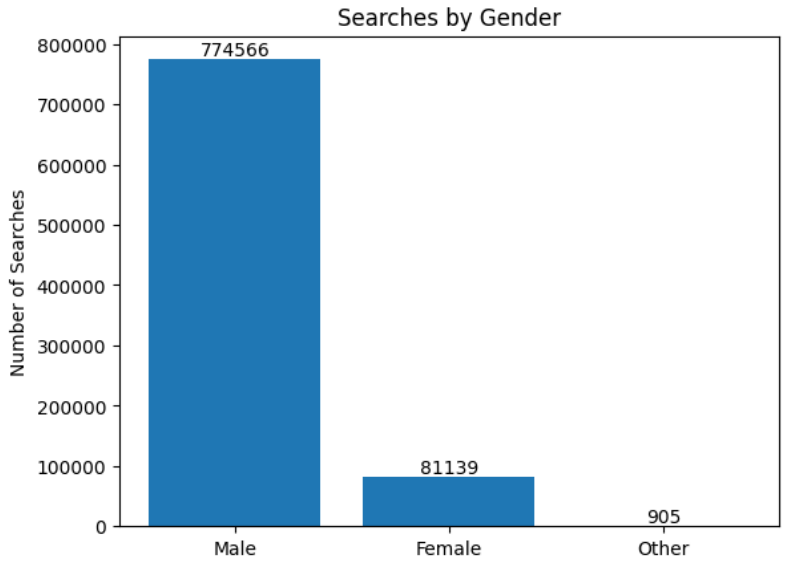

Figure 2.3

2.1.4 Officer-defined ethnicity

The following graph in figure 2.4 demonstrates the number of searches conducted in each officer-defined ethnicity. The majority of the searches are conducted on individuals who present as White, followed by those who present as Black and Asian. The numbers for the "Other" and "Mixed" categories are relatively low in comparison, which suggests that there is a significant discrepancy in the distribution of searches among different ethnic groups.

The higher percentage of White individuals in the graph aligns with the demographic composition of the UK, but the disproportionate representation of other ethnicities suggests a deviation from the expected proportional distribution.

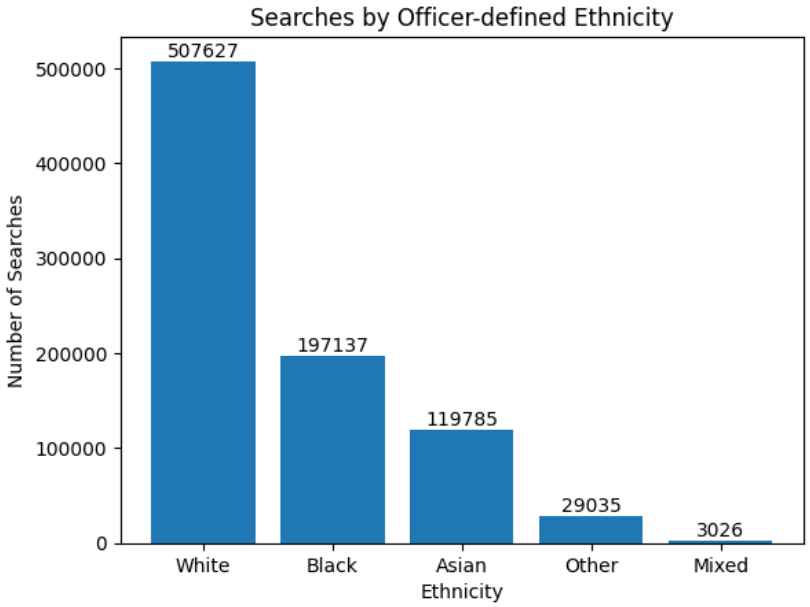

Figure 2.4

2.1.5 Object of search

The "Object of search" is an important variable when analysing police searches. It reveals the reason why the search was conducted and provides insight into the types of criminal activity that law enforcement is targeting.

The data shows a wide range of objects searched for during police operations, from controlled drugs and offensive weapons to stolen goods and articles for use in theft. However, there is a significant variation between the number of searches conducted for certain objects compared to others as seen in figure 5.1, predominantly in controlled drugs.

2.1.6 Outcome

The "Outcome" graph of the dataset lists the possible outcomes of a search. The most common outcome was "No Further Action Disposal," which occurred in the majority of searches. However, there were significant discrepancies between this outcome and others, as we can see on the figure 2.5.

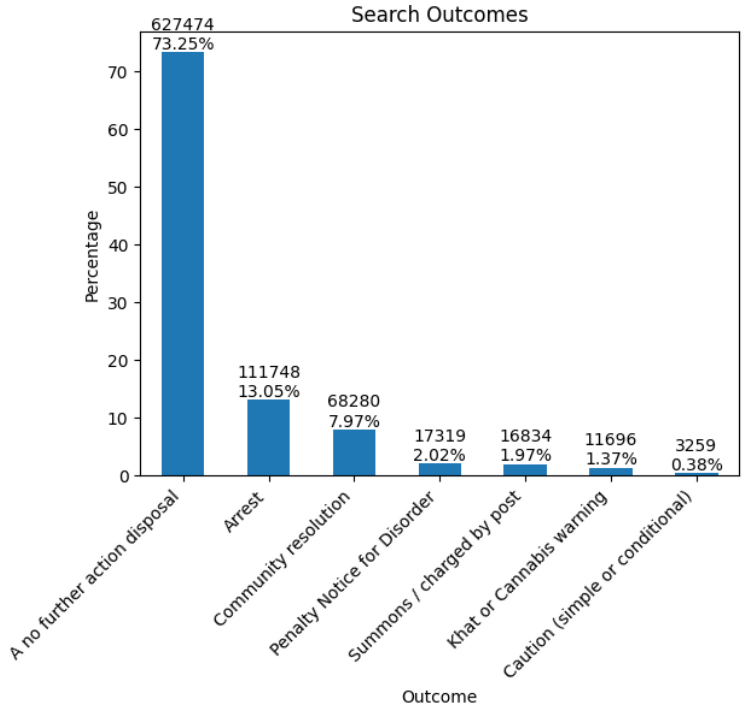

Figure 2.5

2.1.7 Successful Search

Our target variable is if a Search was Successful or not. A search is considered successful if the outcome is positive, that being any outcome different from “No further action disposal”, and related to the search. As shown in figure 2.6 it corresponds to a very unbalanced target in the dataset.

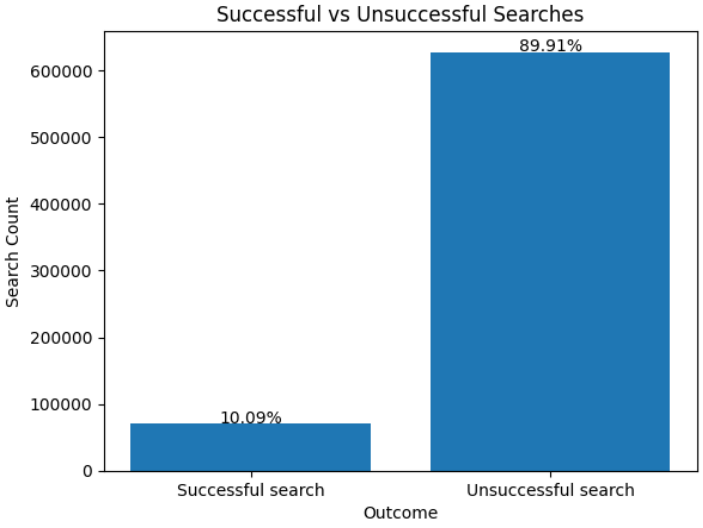

Figure 2.6

2.2 Business questions analysis

In our analysis we have noticed that the stations Leicestershire, Humberside, Lancashire, Metropolitan and West-Midlands have key information missing so we decided to not take them into account. As we can see in the figure 5.6 they have more than 90% information missing on the “Outcome linked to object of search” .

We also put in place a threshold of 30 people, as per recommended by the client, for each gender and ethnicity in each station to remove any data with no significance.

2.2.1 Search for evidence of discrimination between the different ethnicities/genders

After the data analysis, we searched for evidence of discrimination between the different ethnicities. We initially searched for the success rates of searches based on officer defined ethnicities and found that there were no 5% variations between them (figure 5.7). However, when we looked at the data per station, we noticed that most stations had a variation of more than 5% between the success rates of searches for different ethnicities/genders. This indicates that there may be some instances of discrimination happening at specific stations (all the stations are highlighted on the figure 5.8).

We also checked the percentage of ethnicities/genders that are 5% below its station average and we can clearly see a sign of over-search on women, especially minorities. This is also the case of men with the officer-defined ethnicity “Other”. Both situations are seen in figure 5.9.

2.2.2 Verify if women of certain age groups and ethnicities are being asked more often to remove outer clothing

Based on our analysis, we found that some stations had missing data regarding removal of clothing, and therefore, we excluded them from our investigation. These stations were Surrey, North Yorkshire, and Cleveland (figure 5.10)

We compared the successful searches when more than outer clothing was removed between male and female presenting individuals, and the results showed a clear divergence, with almost 10% less success rate on searches conducted on female presenting people (figure 5.11). This suggests that women are being asked more often to remove more than outer clothing than men without any successful outcome after.

When looking at the officer's defined ethnicity, we found that white women were the most searched group by a large margin, as seen in figure 5.12. Thus, the conclusions we could draw were limited due to there only being very few searches on the remaining ethnicities.

Further analysis revealed that there were smaller success rates for White women aged 10 to 17, Asian and Mixed women aged 25 to 34, and White and Black women over 34 as seen in figure 5.13.

Additionally, we concluded that some stations had significant inconsistencies between age ranges and ethnicities, which we could not verify accurately due to the small

dimension of the sample. To ensure data quality, we established a threshold of at least ten people per age range, ethnicity and station to exclude low-value instances, which removed all ethnicities except white.

Figure 5.14 displays the success rates for removing more than outer clothing for white female individuals and figure 5.16 its total numbers by age range and station. Notably, some stations had very low success rates, particularly in the over 35 age range and exhibited significant differences in success rates across different age ranges. All the stations with differences higher than 5% between age-range and success rates below 10% the stations average are marked.

Given the extent of discrepancies and the limited sample size, we could not draw robust conclusions about the efficacy of searches with removal of more than outer clothing for certain genders, age groups and ethnicities in these stations. However, the findings warrant further investigation into the appropriateness and effectiveness of removing more than outer clothing in searches.

2.3 Recommendations

One key area that may merit further investigation is the potential for discrimination against certain ethnicities and genders at specific stations. Our analysis showed that there were significant variations in the success rates of searches for different ethnicities/genders across stations, indicating the possibility of discriminatory practices. As such, we recommend that further investigation be carried out to identify the root causes of these discrepancies and to take corrective action where necessary.

Another factor that may impact our analysis is the difference between self-defined and officer-defined ethnicities (figure 5.16). It is possible that some individuals may identify as a certain ethnicity, but are classified differently by the officer conducting the search. This could introduce inaccuracies into our analysis and could contribute to the discrepancies we observed across stations.

Additionally, our analysis found that women of certain age groups and ethnicities are being asked more often to remove more than outer clothing, which could be indicative of discriminatory practices. To address this issue, we recommend, besides further investigation, that officers receive additional training to ensure that their searches are carried out in a fair and impartial manner, without any discrimination based on age, gender, or ethnicity.

Finally, our analysis was limited by the lack of complete data in some stations. We suggest that all police stations ensure that they collect and maintain accurate and complete data on stop and search practices to enable better analysis and identification of potential issues. In the figure 5.17 we can see what data is missing from each station.

Overall, we believe that our findings and recommendations can help the police to ensure that their stop and search practices are carried out in a fair and non-discriminatory manner, while also improving public confidence in law enforcement.

3. Modelling

3.1 Model specifications

The proposed model is a supervised machine learning algorithm to predict the success of police stop and search procedures in the UK based on various features. The data has been cleaned by dropping rows that belong to certain stations that were previously mentioned.

Firstly, the data was cleaned by removing the stations mentioned in points 2.2 and 2.2.2. Secondly, the missing values of the column “Part of a policing operation” were replaced with False, “Legislation” with unknown and “Outcome linked to object of search” with False. “Outcome linked to object of search” was changed to False in the cases where the “Outcome“ was “A no further action disposal”.

A custom scikit-learn Transformer was created to extract the month, day of the month and week and hour from the date column and create new columns with them. A new column named “Success” was added, which is set to 1 if the outcome of the search is successful, and 0, if it is unsuccessful.

Considered features for the analysis :

Categorical

- Legislation
- Object of search
- Part of a policing operation

Numerical

- Hour
- Day of the Month
- Day of the Week
- Month

To handle categorical features, we use one-hot encoding after imputing missing values.

A train-test split was performed, reserving approximately 30% of the data for evaluation purposes. These tests are described in Section 3.2.

We ended up using the LogisticRegression as it presented the best overall precision and recall. The following parameters were chosen:

- C=1
- class_weight='balanced'

To make a final prediction, if the predicted probability is equal to or higher than 47%, the observation is predicted as True, and if it is lower, it is predicted as False. This threshold value was determined by optimising for both precision and recall using a custom function.

3.2 Model performance and expected outcomes

After analysing the preliminary results of our model prototype for predicting the success of a police search what is expected of our model is :

- F1 score: 0.35
- Recall score: 0.87
- Precision score: 0.22

The F1 score, precision, and recall values indicate the balance between accuracy and completeness of the model's predictions. They provide insights into the overall performance of the model in identifying positive cases and minimising false positives.

Our model achieved a relatively low F1 score, indicating that it may struggle to accurately predict the success of police searches. However, it also achieved a relatively high recall score of 0.87, suggesting that it is able to identify a significant portion of successful searches. The precision score was low at 0.22, which implies that the model tends to generate a large number of false positives, figure 5.18 shows the model's confusion matrix.

Additionally, we used a function to test if our model exhibited discrimination against certain sensitive groups (based on gender, ethnicity and age range). We divided the police stations in most and least biased according to their precision rate. The former refers to a precision rate below 5% and the latter to one above 5%. The results of this test are displayed on table 3.1.

Table number 3.1

	Number of most biased stations	Number of least biased stations	Average Difference
Gender	27	9	0.14
Ethnicity	34	2	0.17
Age Range	34	1	0.24

As shown on table 3.1, our model failed to meet our requirements for both gender, ethnicity and age range, due to the average differences of precision scores between protected groups being above 0,05 (5%).

Moving forward, we plan to address the bias identified in our model by exploring alternative algorithms and preprocessing techniques. We also plan to conduct a more thorough analysis of the data to identify potential confounding variables and to develop more nuanced performance metrics. Finally, we will work with our client to incorporate these insights into their existing workflows and to develop tools for ongoing monitoring and evaluation.Our model prototype shows promise for predicting the success of police searches, but also exhibits significant bias against certain sensitive groups. We will continue to refine and improve our model in order to meet our client's requirements for accuracy, fairness, and accountability.

3.3 Alternatives considered

During the development process, we tested several models and features, including Logistic Regression, Random Forest, and LGBM classifiers. Although LGBM and Random Forest models showed very similar results, they exhibited worse recall compared to the LogisticRegression, which was ultimately chosen as the final model based on its trade-off between precision and recall.

We also attempted to improve the model's performance by incorporating additional features such as latitude and longitude, and by addressing class imbalance through undersampling techniques on the training data. However, these attempts did not result in significant improvements.

In an effort to address issues of fairness and bias, we also trained the model using sensitive features such as gender, officer-defined ethnicity, and age range. However, this led to worse performance, lower precision and an increase in the number of most biassed stations. Therefore, we ultimately decided to keep sensitive features on the final model.

Table number 3.2

	F1 score	Recall	Precision	Average difference (age, gender , ethnicity)
LogisticRegression	0.35	0.87	0.22	0.18
RandomForestClassifier	0.35	0.84	0.22	0.19
LGBMClassifier	0.35	0.85	0.22	0.19
LogisticRegression (with undersample)	0.35	0.84	0.22	0.19
RandomForestClassifier (with undersample)	0.35	0.80	0.21	0.18
LGBMClassifier (with undersample)	0.35	0.85	0.22	0.19
LogisticRegression (with sensitive info)	0.35	0.79	0.22	0.19
RandomForestClassifier (with sensitive info)	0.35	0.81	0.23	0.18
LGBMClassifier (with sensitive info)	0.37	0.78	0.24	0.20
LogisticRegression (with Latitude and Longitude)	0.35	0.85	0.22	0.19

4. Model Deployment

4.1. Deployment specifications

To replicate the model deployment, you would need to start by importing the necessary files, including columns.json, dtypes.pickle, and pipeline.pickle. The model was then integrated into a Flask web server and deployed on a railway. The app's database uses PostgreSQL.

The app's database is a PostgreSQL database hosted on Railway, just like the API. Within the database, there is a table called "Prediction" that consists of four fields: observation_id, observation, predicted_outcome, and outcome. The "observation_id" field is a unique text field serving as the primary key. The "observation" field is a text field used to store the details of each observation. The "predicted_outcome" and "outcome" fields are boolean fields that store the predicted and true outcomes of the search, respectively.

To integrate with the database, the app uses the Peewee ORM (Object-Relational Mapper). The connection to the database is established using the DATABASE_URL environment variable. The structure of the "Prediction" table in the database is defined in a class called "Prediction," which extends the Peewee Model class. This class allows the app to perform CRUD operations (Create, Read, Update, Delete) on the records in the "Prediction" table.

Since a custom transformer was created to handle feature engineering, it needs to be imported. The API also handles all data cleaning.

The API includes two endpoints.

1. **Should_search.**
    - This endpoint receives a request containing officer-inputted information and returns a prediction as to whether a person should be searched or not. It is expected to receive the following content with the respective types:
        - "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>

- The endpoint only considers the features mentioned in section 3.1 to make a prediction, and it is designed to be able to accept new categories and variables.

- If the observation_id and Object of Search is missing from the request, an error will be returned. To address missing values for certain fields, such as Part of a Policing Operation and Legislation, the API has a function named "fill_missing_categorical_columns" that will automatically fill these values with False and “unknown” the same way as It’s done on the data cleaning process. If the Part of a Policing Operation column is present, it will be converted to a boolean type.

- The Latitude and Longitude values in the request will only be accepted if they fall within the expected range of the UK.

- In the case of the Date field, the API has a function named "check_date_format" that will check if the Date is in the proper format. If it is not, the function will attempt to parse the date string using different format strings until it succeeds. If the Date is missing important values such as hour, day, or month.

- If the requirements are not met an error 405 will be raised with a message regarding the specific problem .

2. **Search_result.**

    - This endpoint is a POST request that updates the outcome of a search prediction based on the provided observation ID. It expects a JSON request containing the observation ID with the type string and the new outcome, a boolean value. If the observation ID exists in the database, the prediction outcome is updated and a JSON response is returned containing the updated prediction's details. If the observation ID does not exist, a 405 error is returned with a message indicating that the provided observation ID does not exist.

4.2. Known issues and risks

The implemented API is designed to be flexible and able to accept new categories and values. This allows for future expansion of the model to consider additional features and data points, improving the accuracy of the predictions. However, this flexibility also poses a risk as it increases the potential for errors in the input data. To mitigate this risk, thorough validation and cleaning of the input data is essential, as well as ongoing monitoring of the performance of the model as new categories and values are added.

One known issue with the current implementation is the use of railway data to determine the proximity of a search to a railway station. This data may not always be accurate or up-to-date, potentially leading to errors in the prediction. It is important to regularly review the data sources used by the model and ensure that they are reliable and appropriate for the intended use case.

Another potential risk with the mentioned implementation is the handling of personal information. The API collects data such as the individual's ethnicity and the reason for the search, which could be considered sensitive information. Proper measures must be taken to ensure that this information is handled securely and in compliance with applicable data protection laws and regulations.

Using a free tier cloud server for hosting the API and database introduces certain risks due to the limited available resources. The allocated 512 MB of RAM may prove insufficient for handling larger workloads or spikes in traffic, potentially resulting in slower response times or even system crashes. Similarly, the shared CPU or container may limit the processing power available, further impacting the overall performance of the system. Additionally, the 1 GB disk space constraint may restrict the amount of data that can be stored and processed, potentially leading to data management challenges. It is important to carefully assess the resource limitations and evaluate whether they align with the requirements of the application and expected usage patterns.

Furthermore, it is essential to consider the usage limit associated with the free tier. The budget of 5 € or 500 hours of usage implies that once the allocated limit is exhausted, the services may be temporarily suspended or inaccessible until the limit is reset or upgraded. This can pose a significant risk for critical applications that require continuous availability and uninterrupted service. Evaluating the expected usage patterns and comparing them against the allocated budget is necessary to ensure that the free tier can adequately support the needs of the application. Upgrading to a paid tier with more substantial resources and higher usage limits may be a viable solution to mitigate the risks associated with limited resources and usage constraints.

In summary, while the implemented model and API offer valuable tools for improving the accuracy and efficiency of police searches, there are potential risks and issues that must be considered and addressed. Ongoing monitoring, validation, and review of the data sources and handling of personal information are critical to ensure the reliability and ethical use of the API.

5. Annexes

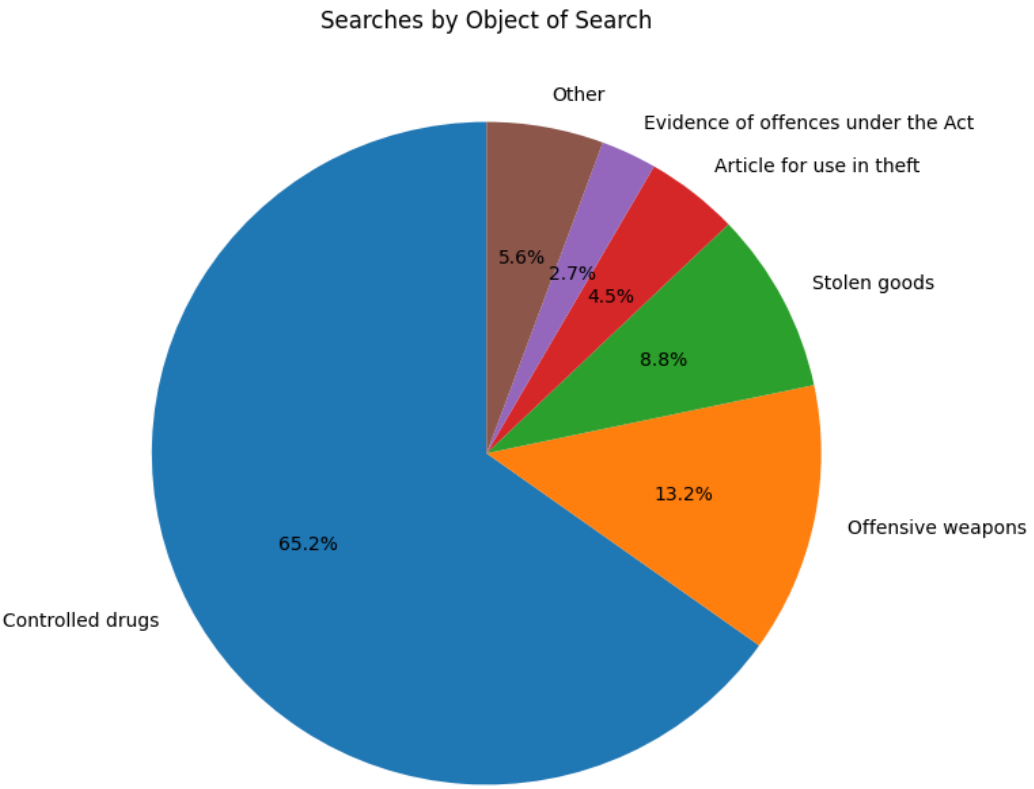

Figure 5.1

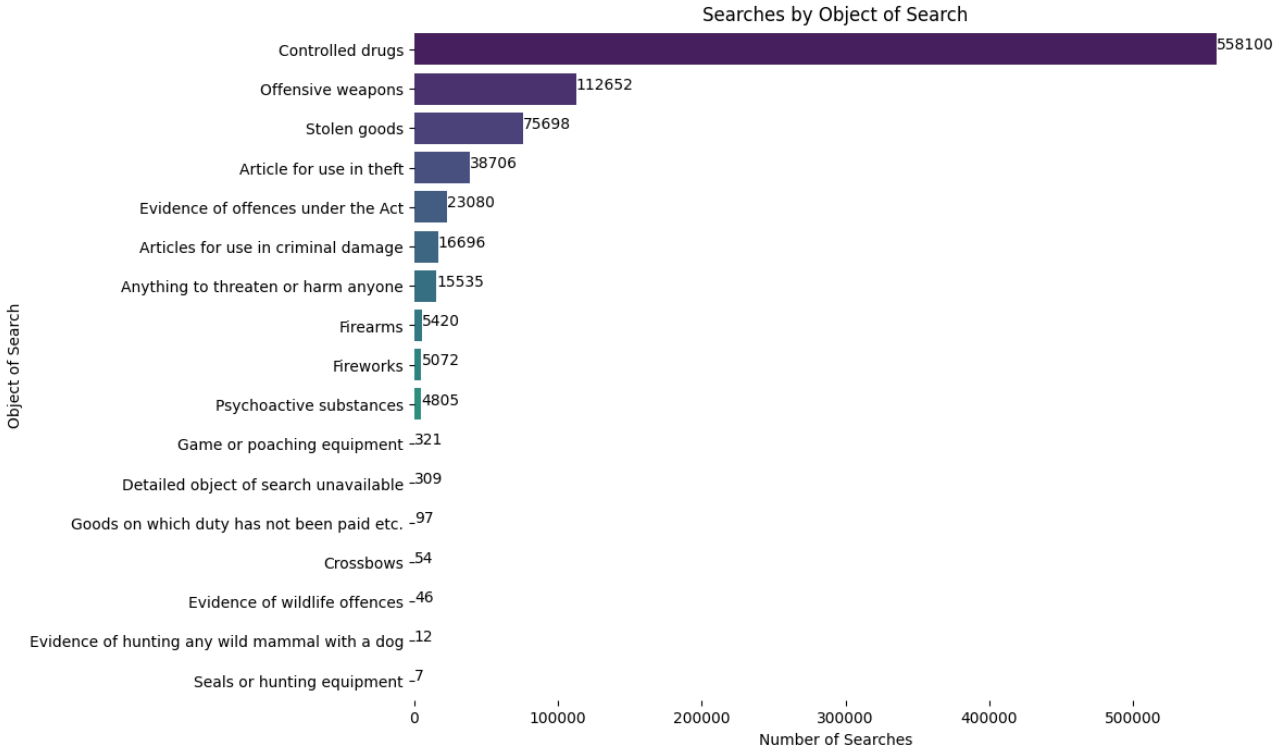

Figure 5.2

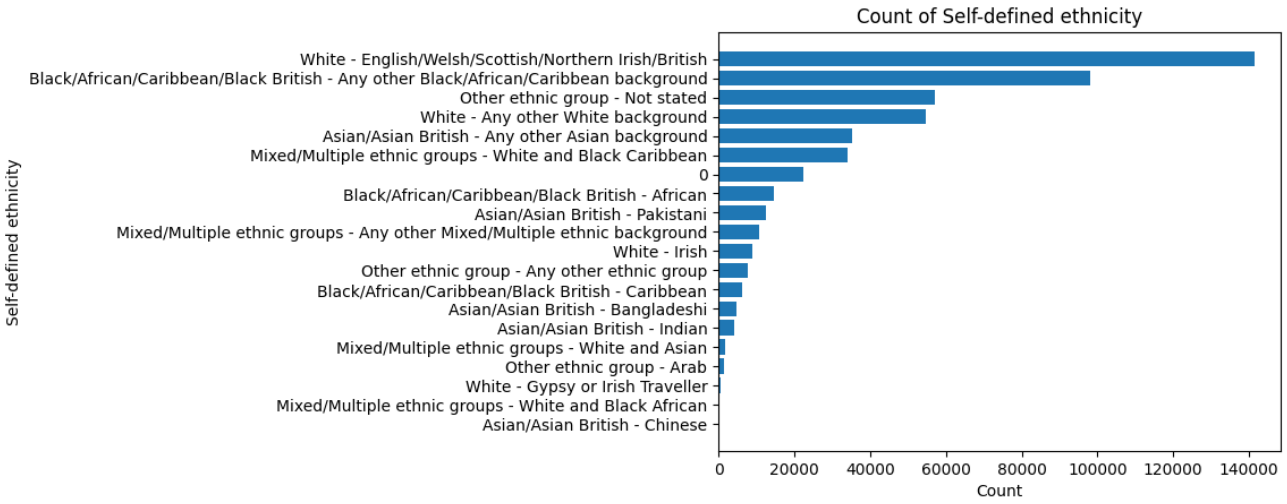

Figure 5.3

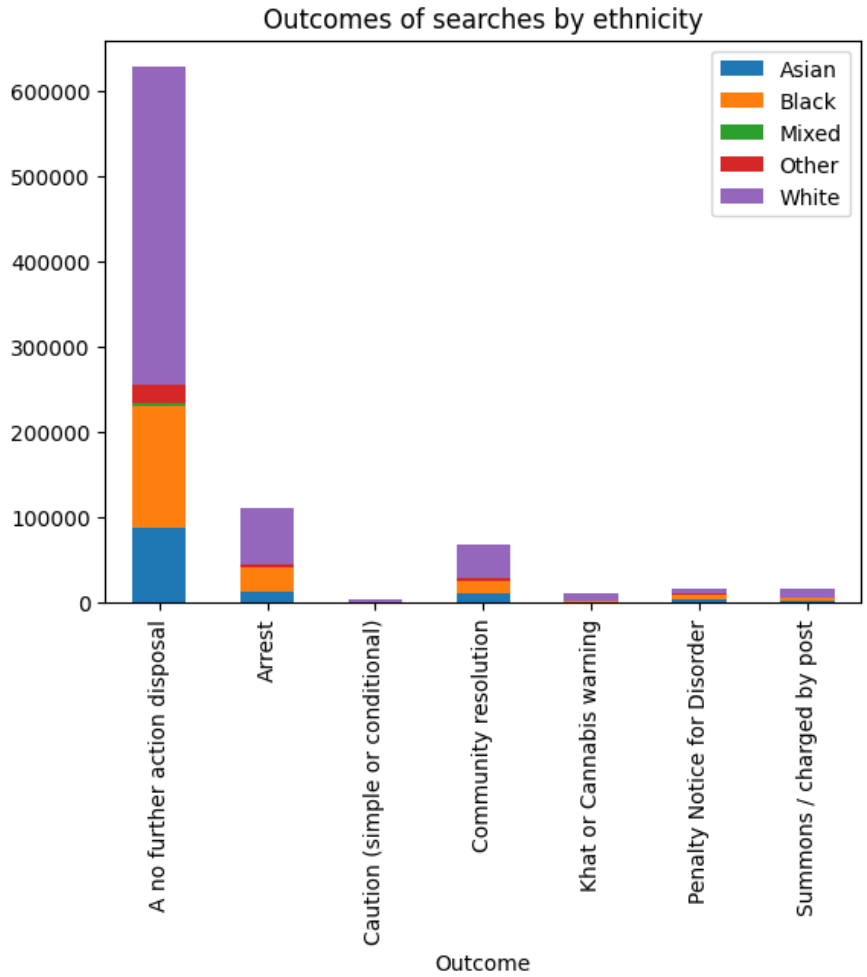

Figure 5.4

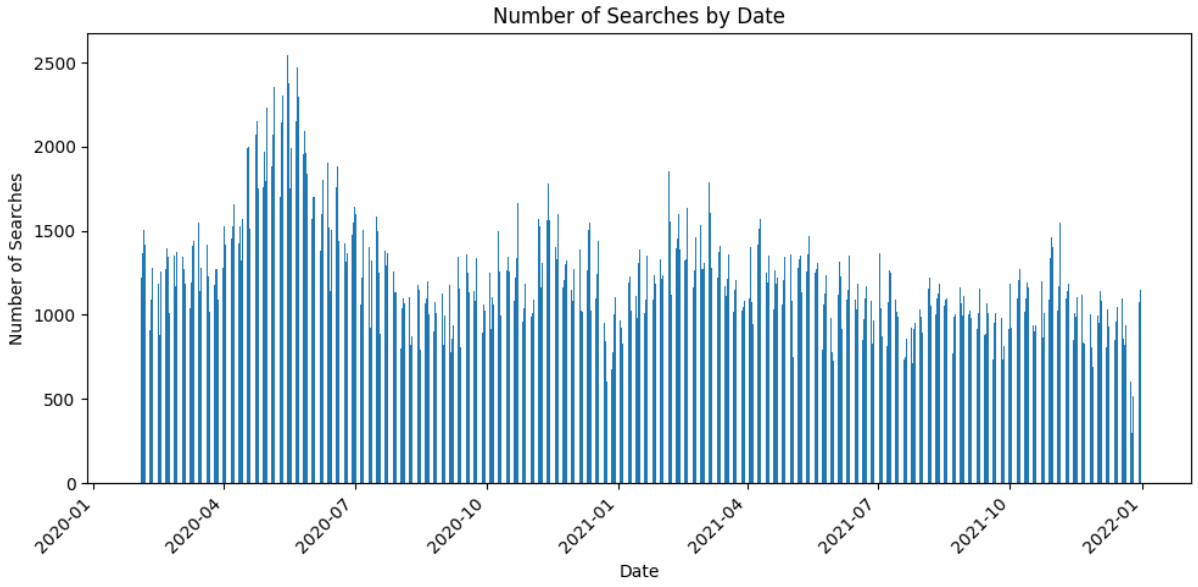

Figure 5.5

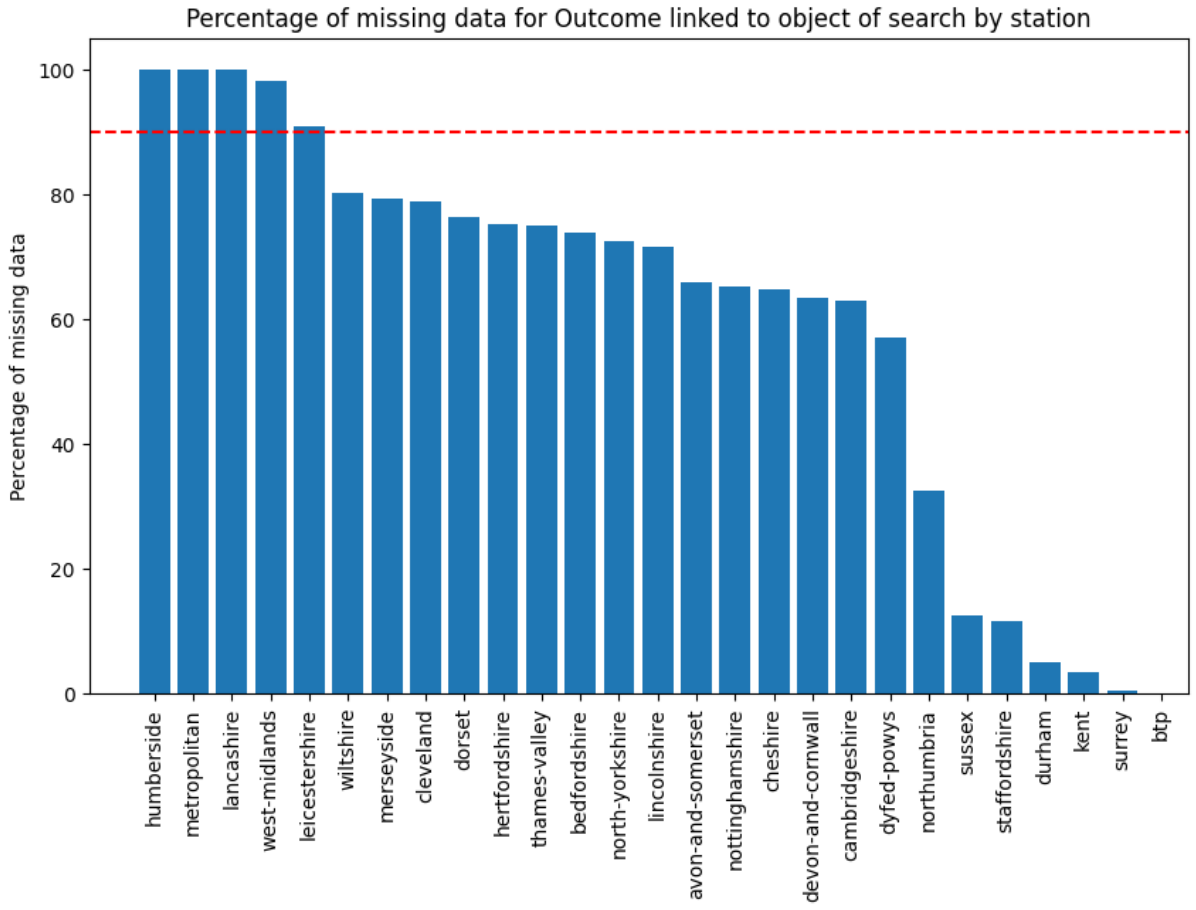

Figure 5.6

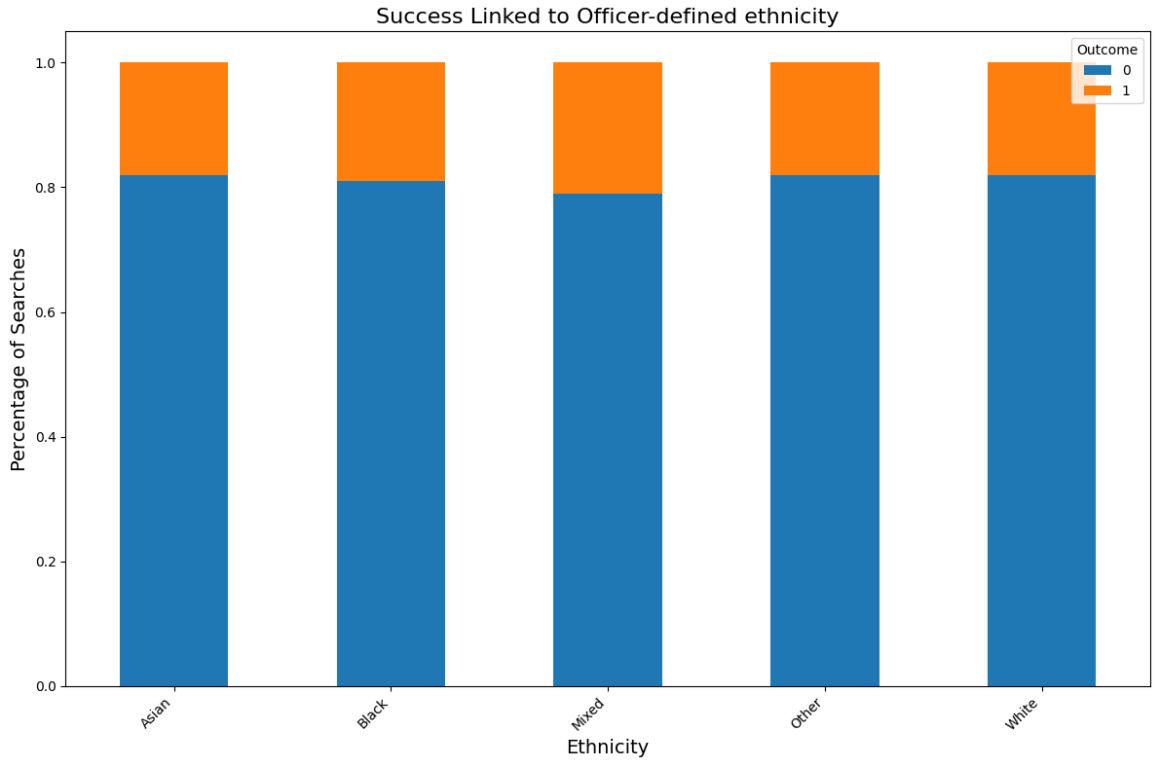

Figure 5.7

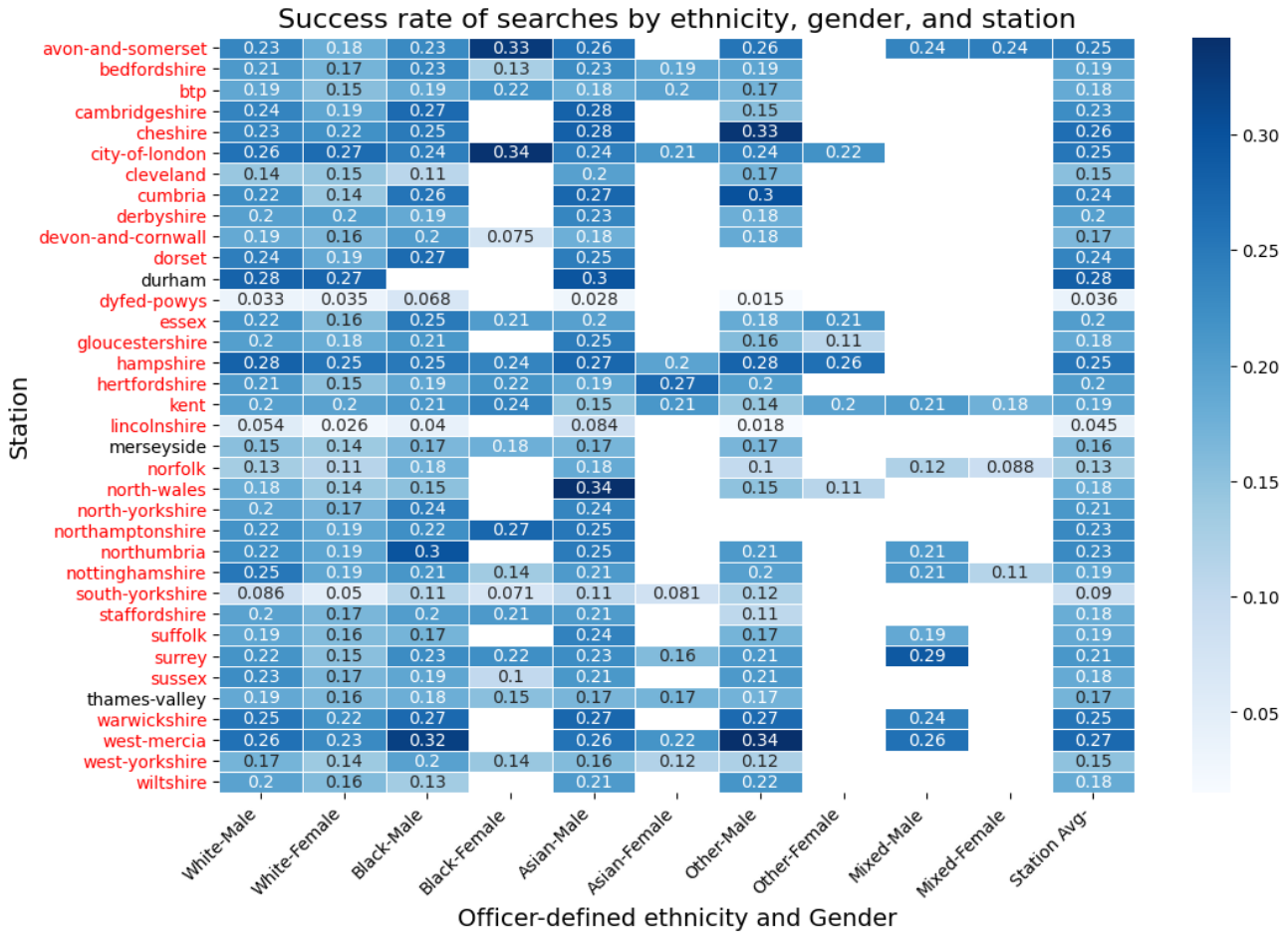

Figure 5.8

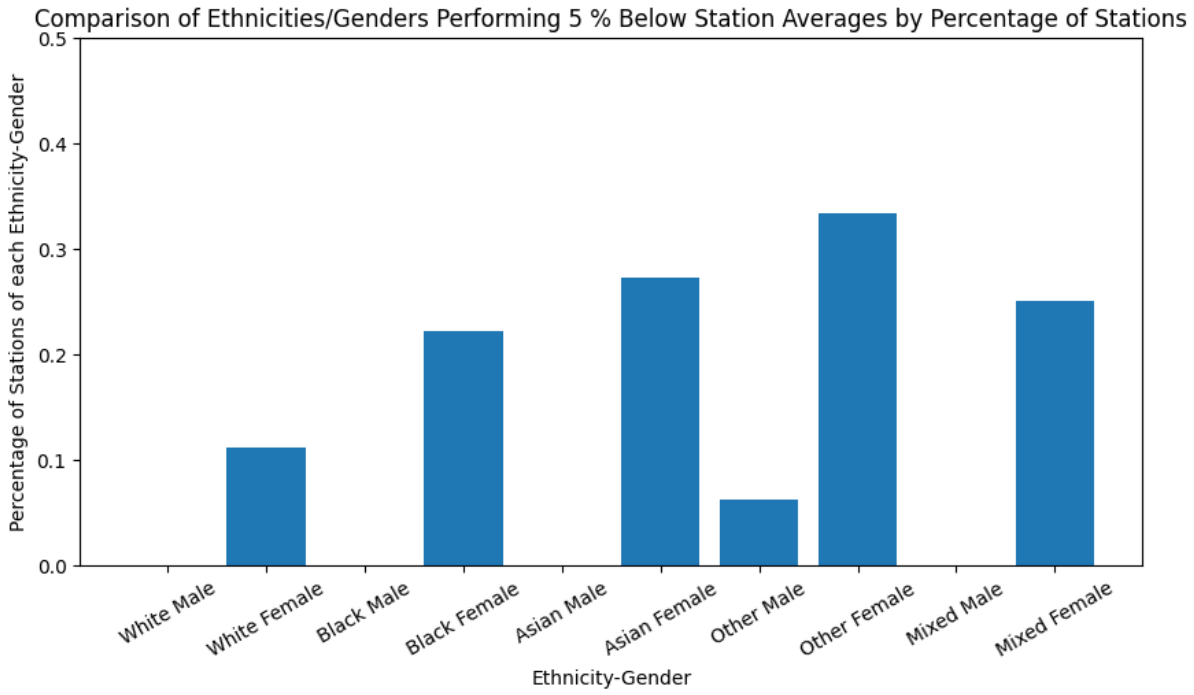

Figure 5.9

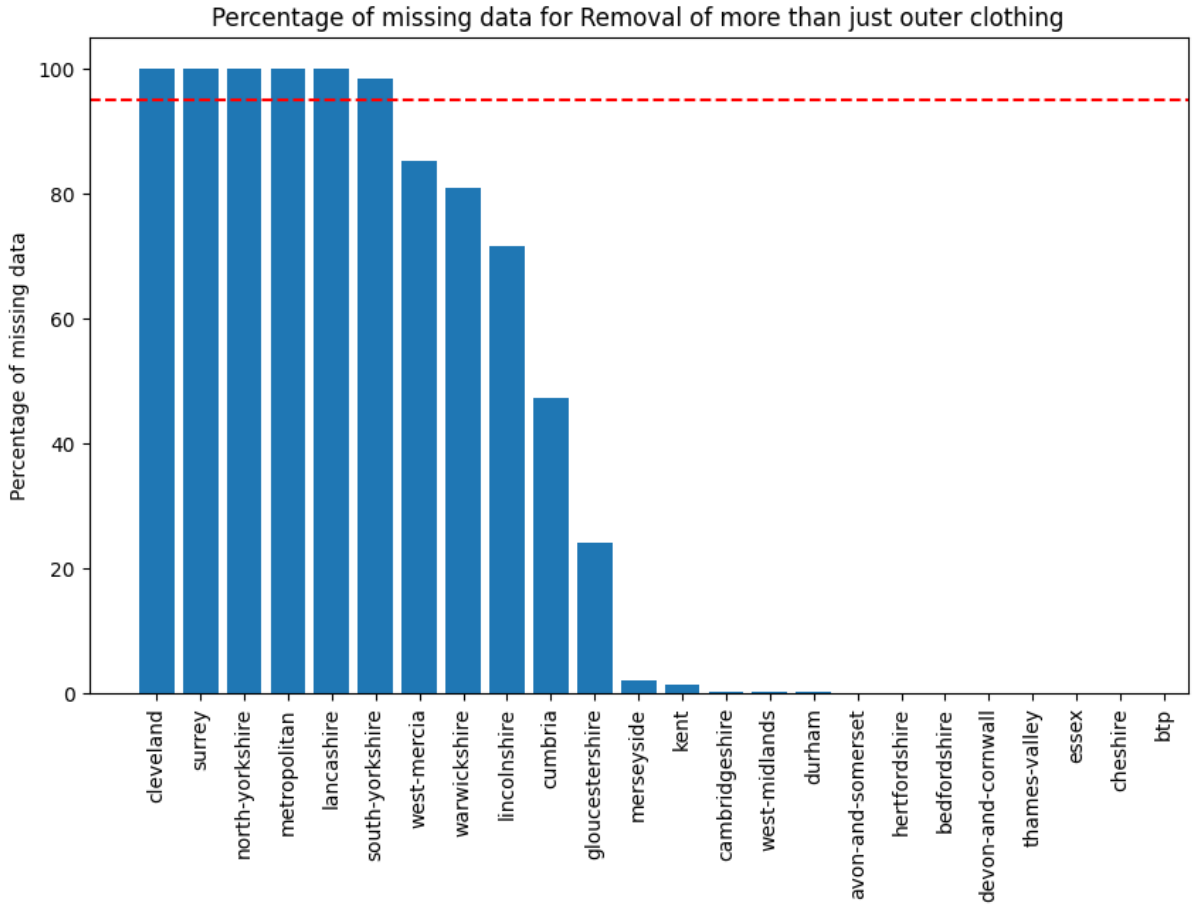

Figure 5.10

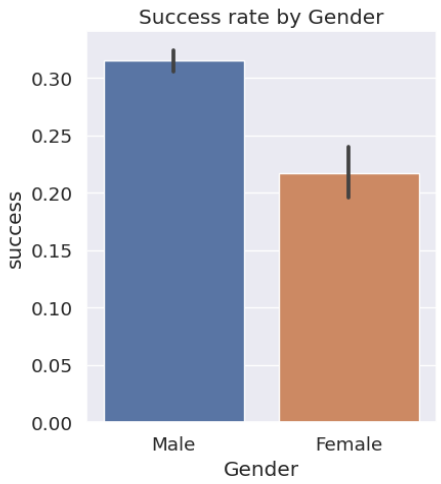

Figure 5.11

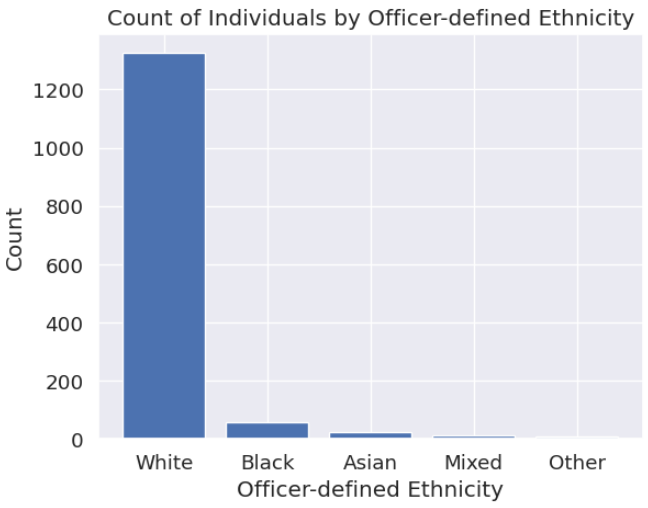

Figure 5.12

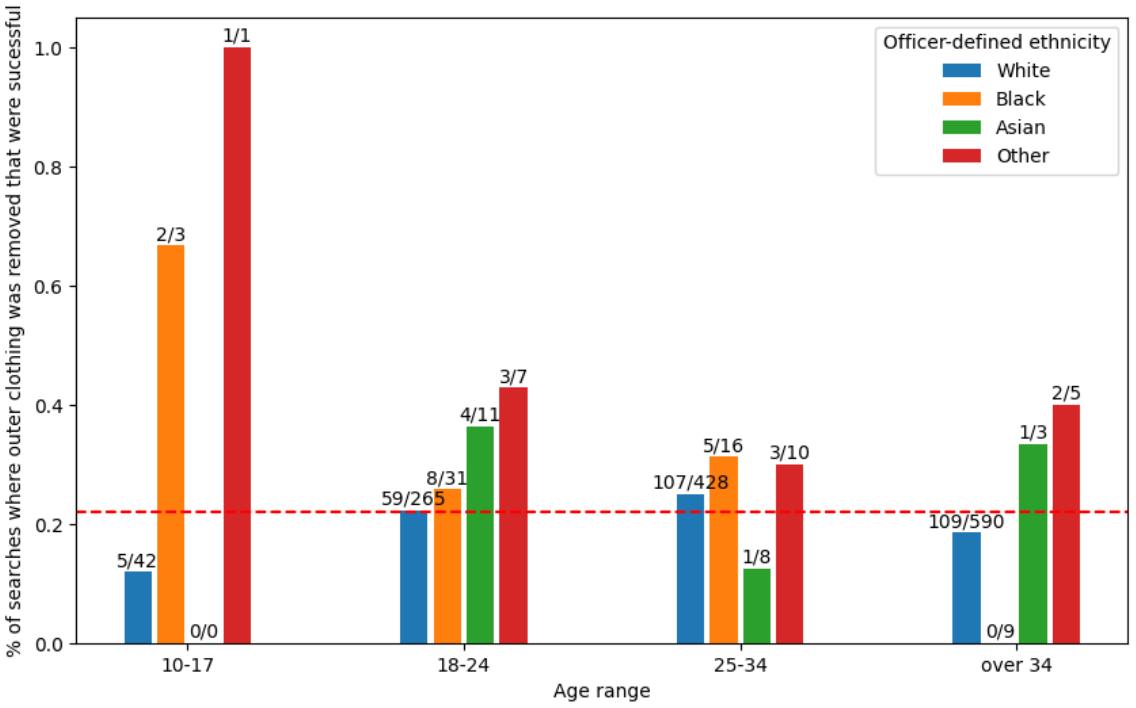

Figure 5.13


Success rate of searches by age range
avon-and-somerset 0.15 0.24 0.29
cheshire 0.38 0.38
cumbria 0 0.2
derbyshire 0.36 0.15
devon-and-cornwall 0.13 0.098
dorset 0.14
hampshire 0.25 0.31 0.25
hertfordshire 0.32 0.28
kent 0.36 0.4 0.43
merseyside 0.33 0.26 0.24
norfolk 0 0.091 0.071
north-wales 0.4
northumbria 0.091 0.3 0.037
south-yorkshire 0.062
staffordshire 0.3 0.38 0.11
suffolk 0.23 0.061
sussex 0.18 0.17 0.35
warwickshire 0.33 0.23
west-mercia 0.36 0.17
west-yorkshire 0.29 0.045
Station
18-24
25-34
over 34
Age range
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

Figure 5.14

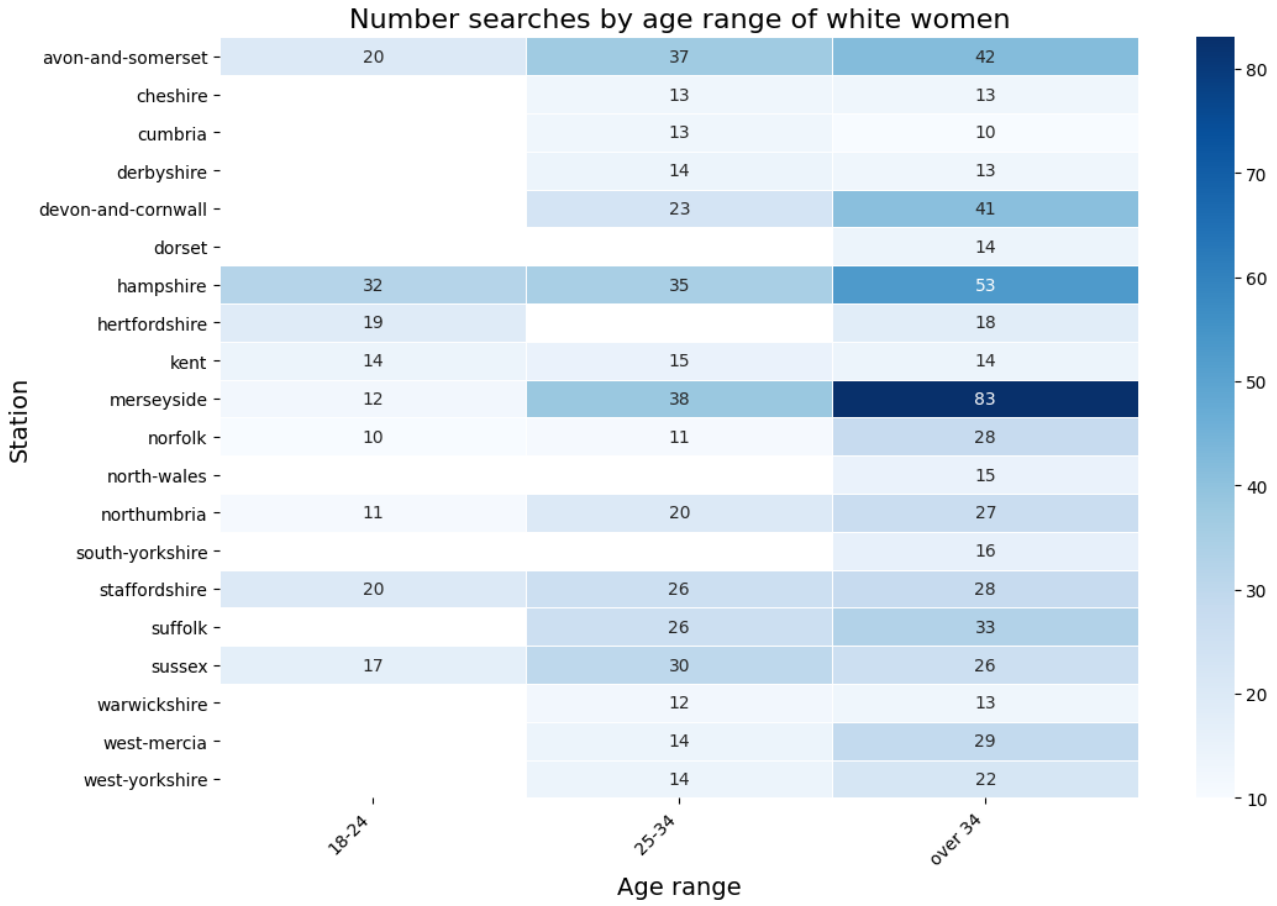

Figure 5.15

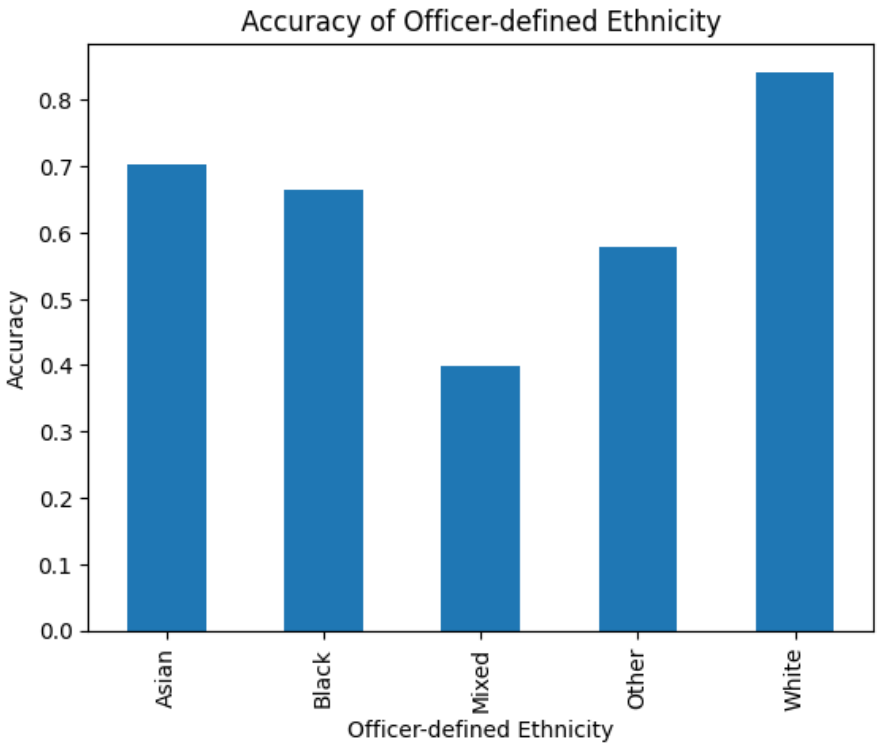

Figure 5.16

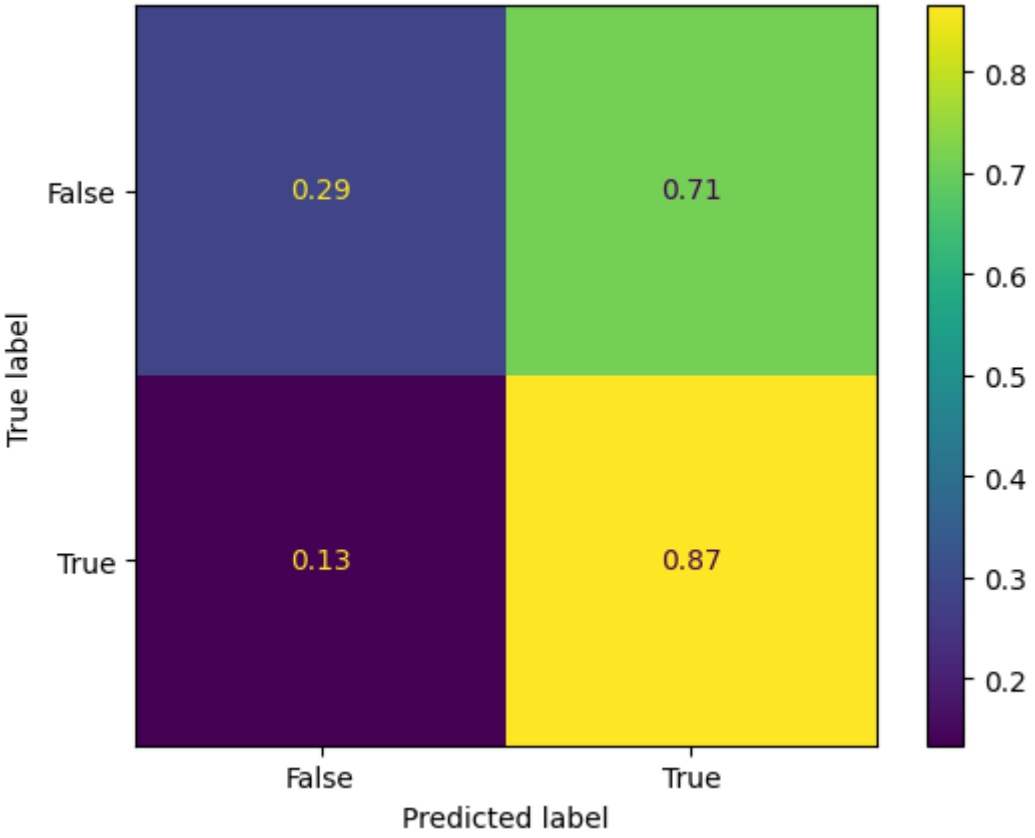
0.29
0.71
0.13
0.87
False
True
True label
False
True
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Figure 5.18

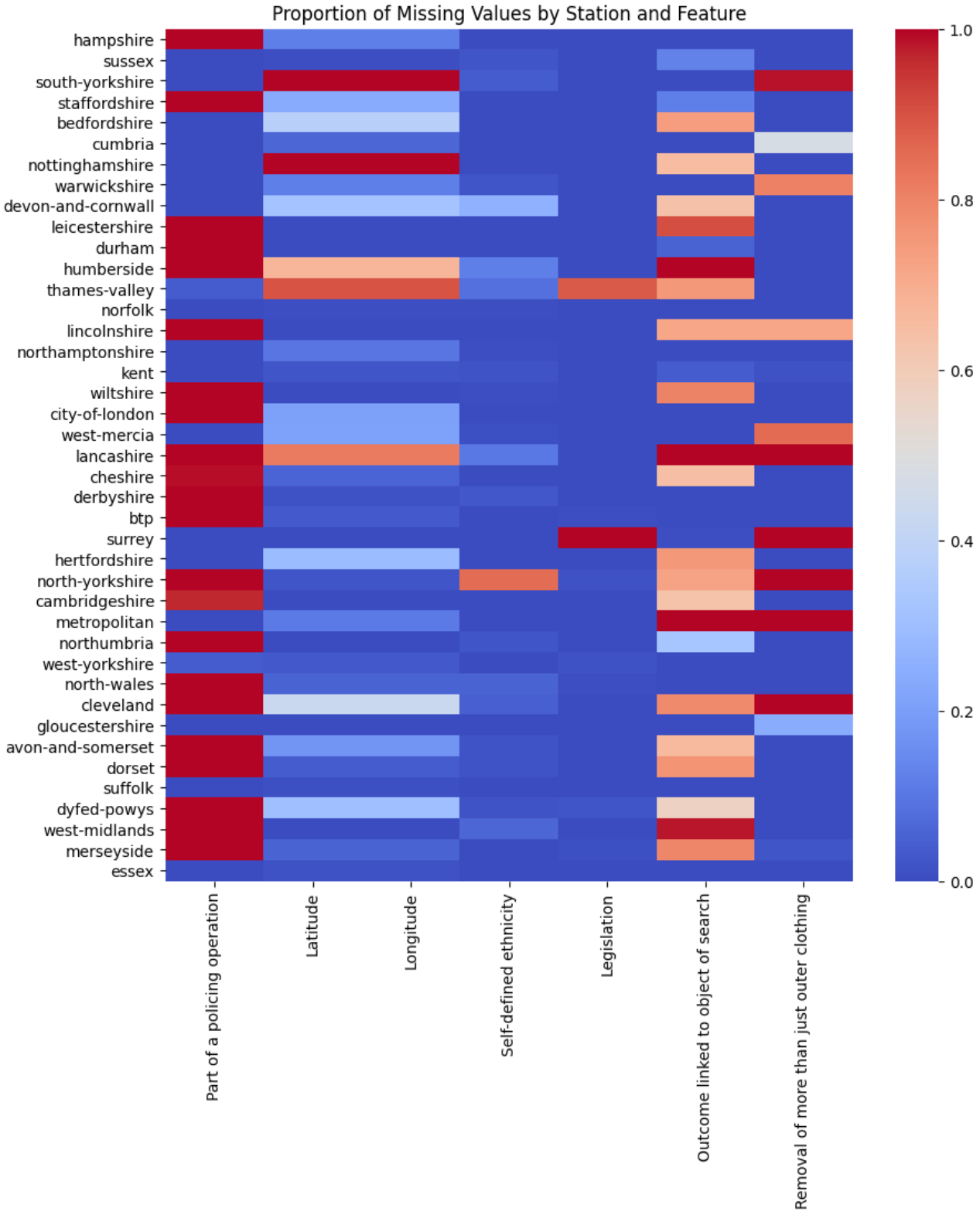
Proportion of Missing Values by Station and Feature
hampshire
sussex
south-yorkshire
staffordshire
bedfordshire
cumbria
nottinghamshire
warwickshire
devon-and-cornwall
leicestershire
durham
humberside
thames-valley
norfolk
lincolnshire
northamptonshire
kent
wiltshire
city-of-london
west-mercia
lancashire
cheshire
derbyshire
btp
surrey
hertfordshire
north-yorkshire
cambridgeshire
metropolitan
northumbria
west-yorkshire
north-wales
cleveland
gloucestershire
avon-and-somerset
dorset
suffolk
dyfed-powys
west-midlands
merseyside
essex
Part of a policing operation
Latitude
Longitude
Self-defined ethnicity
Legislation
Outcome linked to object of search
Removal of more than just outer clothing
1.0
0.8
0.6
0.4
0.2
0.0

Figure 5.17