NHS Data Analytics Report

Business problem – To reframe the business problem I propose to investigate 'How are NHS resources and staff being used to meet the demand for its services.'

This business problem will help us share some insights into the current use of resources , how staff is currently handling the workload, and the possible relationship between cancelled appointments.

Industry context – The healthcare system in the UK, the current lack of staff, the funding crisis and more than appointments 22,000 being cancelled[1]. By setting the context of the current NHS crisis, we can explore how current resources are being used and how to utilise them more efficiently by taking a data-informed approach.

Business objectives

1. To decrease cancelled appointments by finding possible explanations and solutions as to why individuals cancel appointments or do not attend last minute.
2. To decrease services that are classified as 'unknown'. This prevents the NHS to understand clearly how services are being deployed.
3. To increase staff capacity in locations with the highest demand for appointments. And so, locations with high demand will be more manageable.
4. To understand NHS's utilisation capacity.

For instance, a question I would like to pose for the rest of the Data Analyst is if is there a relanthshi0 between lower-income families and missed appointments. This could be due to not being able to ask permission at their job or to caring responsibilities.

Analytical approach

To refer to the code snippets, libraries, functions, and methods used for this piece of work please refer to 'Figure 1', where I state each piece of code for the assignment. I have developed various processes to complete EDA along with using the metadata to understand the origin of the datasets. Firstly, I explored each individual dataset to explore its columns, and how they differ in terms of the number of records and shared columns between the datasets[2]. For example, the ad and ar dataset show a large jump from Q2 to Q3, this was found using the described method. After further EDA I found the reason for this was derived from outliners and very large value difference between categories. For further detail please refer to the code snippet findings on the Jupyter notebook.

I will proceed to answer some questions proposed to investigate our business objective. (1) How many locations are there in the data set? Both ad and nc show the same number of locations 106 unique locations. (2) 'What are the five locations with the highest number of records? ad and nc both share Kent and Medway as their second most popular location, and Northwest London is shared by both datasets within their top five locations. (3) How many service settings, context types, and national categories are there? For nc the most popular service setting is General Practice; the context type is Care Related Encounter, and the national category is inconsistent mapping (3.1) How many appointments status is there? for ar, Attend is the highest with 2.32.137 and Unknown with 201.324. The overall feeling from this exploration is the overwhelming number of categories such as other, unmapped and unknown is very high. I would like to pose the following question to the data team -

---

[1] Please refer to Quettevelle, H. (2022). Inside the staffing crisis that's crippling the NHS. *The Guardian.* Available at: [Inside the staffing crisis that's crippling the NHS (msn.com)](). [Accessed 19/10/22]

[2] I will be using 'ad' to refer to actual_duration.csv, 'ar' as appintments_regional.csv and 'nc' to refer to national_categories.xlsx.

‘Why is there such a high count of categories such as unmapped, other, and inconsistent mapping through the datasets?. (4) Between what dates were appointments scheduled? For nc and ad , the min and max were 2021-12-01 and 22-06-20, so the data was taken exactly for 6 months and 29 days. (5) Which service setting was the most popular for NHS Northwest London from 1 January to 1 June 2022? General Practice with 2.104 and Other with 1.318. (6) Which month had the highest number of appointments? For nc 2021 -11 with 30.405.070 counts and for ar is 22-03 with 27’170.002. Here, we can see both datasets have a different month with the highest number of appointments. (7) What was the total number of records per month? For ar the highest record was 22-03 with 21,236 and nc was 22-03 with 83.922 records. This is key information as both datasets share the same date for the highest number of records. It would be meaningful to explore What was the reason behind that with the data team. For further detail on the code snippet and on each process undertaken to answer each question, please refer to the code snippet findings on the Jupyter notebook and to Figure 1 and Figure 2.

Visualisation and insights

For the seasonal trends and for the visualisation by using line plot. When exploring the nc data frame and service settings, national categories and its different categories by using hue argument. The different categories are too many and defer on the scale. Thus, when using a line plot, the audience might be unable to understand and spot trends (Refer to Figure 5). Hence, I have decided to group categories. Into sub-groups based on similar scales or individual categories. By doing this it has been easier to stop seasonal trends by visualization. For example, by subsetting only to GP, we can spot clear trends there is a peak in 201-10 decreasing steadily until 2021-12. (Refer to Figure 6,7 for individual graphs). Due to space constraints, I cannot expand on detail on each service setting or context type. However, some general findings are that both the Service setting and National categories, share the same seasonal trends as follows; 2021-10 high peak for the service setting then it decreases in 2021-12 and there is steady growth after 2022. Please refer to individual findings from each graph on the Jupyter notebook.

To explore the most specific season trends per month, I have used bar plots for clearer visualisation as a line plot did not seem the best choice for the scale of data. (Please refer to figure 8). Some interesting findings are that general practice and unmapped remain the most popular category throughout the seasons. Please refer to individual findings from each graph on the Jupyter notebook[3].

Patterns and predictions

To answer questions in regard to staff and capacity in the network. I have created average and percentage columns with the ar data frame. The utilisation of resources per month was 1’013.502 compared to the max utilization of 1’200.000 which leaves space for greater utilization (Please refer to Figure 4). To explore capacity utilization per month refer to figure 10, a bar plot which displays the months with the highest utilization. From here we can conclude 2021-10 and 11 and 2021 03 have the highest utilization capacity. Some general findings regarding healthcare professionals attended visits and appointment mode increased in 2021-11 and increase again in 2022-03 (Refer to figures 11, 12,13). Lastly, the boxplot showed the issues with the data, the distribution of the outliers and the problem with very large scales on the y-axis. (Refer to Figure 14). To plot better the changes om service setting over time I have used a bar plot (Refer to Figure 15) where we can infer, 2021 had the highest appointment count compared to 2021 and the highest month was 2021-11.

---

[3] Please refer to Figure 3 and 8. To explore the most popular hashtags.

- Has there been adequate staff and capacity in the networks?
- What was the actual utilisation of resources?

To conclude, some questions were not possible to answer. For example, It was impossible to determine if the NHS should increase staff levels based on the utilization capacity. There is little information regarding staff. As Some of the data limitations found have prevented this investigation from suggesting recommendations for specific business objectives. I will expand on this further in the presentation. Lastly, according to the calculation the NHS does not use its maximum capacity per month but from this piece of data we cannot infer if the utilization of resources is adequate or not.

Data Limitations and suggestions

In order to address our business problem and to investigate the utilisation of NHS resources, it would have been to have data on the staff and how it copes with the demand. Moreover, it is unclear why datasets with the same columns have a different number of records, this causes doubts about the validity of the data.

The greatest limitation found in the data was the problem of distribution and extreme value difference on a time series analysis as was the case for this NHS analysis. Throughout the analysis, there have been various issues when creating visualisations. For instance, notice how some of the lowest counts for the total number of appointments are 0-1 and the highest 30'000.000. I have explored the possible solutions in detail in the Jupyter notebook[4]. This will be explored further with the data team.

.

[4] Please refer to Cerliani, M (2021). *Anomaly Detection with Extreme Value Analysis.* Available at: [Anomaly Detection with Extreme Value Analysis | by Marco Cerliani | Towards Data Science]. [Accessed 19/10/22].

Appendix

Figure 1 - Code, functions, methods and libraries imported for Data Analysis

Note: Some of the code used for different assignments is repeated throughout the Jupyter Notebook and hence, is not written for each Assignment. Only the code unique to each assignment is written per row.

Assignment Week	Use of Github repository
Assignment 2	**Libraries imported:** pandas, numpy **Code snippet explanation: pd.read_csv** describe(), info(), isnull(), sum(), head(), info() use sub setting techniques to answer questions where data had to filter and manipulated, nunique(), apply(), value_counts(), to_frame()
Assignment 3	**Libraries imported:** datetime **Code snippet explanation:** dtypes, astype, agg(['min', 'max']), df.loc & to subset data with conditions, groupby(), sum() dt.year, dt.month, sort_values(by=, ascending = True)
Assignment 4	**Libraries imported:** seaborn, matplot **Code snippet explanation:** sns. Lineplot, plt.title plt.xlabel, plt.ylabel, reset_index(), sns.barplot
Assignment 5	**Libraries imported:** plotly.express **Code snippet explanation:** for loop with an if condition, pd.Series, df.rename, pd.DataFrame, x = pl.bar
Assignment 6	**Code snippet explanation:** round, /, *, sns.boxplot, range, df.shape, plt.hist,

Figure 2 – General Findings

Questions	Actual Duration (ad)	Appointments Regional (ar)	National Categories (nc)
What is the number of locations?	106		106
What are the five locations with the highest number of records?	1. North Norfolk and Waveney 2. Kent and Medway 3. North West London 4. Bedfordshire Luton and Milton Keynes 5. Greater Manchester		1. North West London 2. Kent and Medway 3. Devon 4. Hampshire and Isle of Wight 5. North East London
What is the number of service settings?			General Practice 359274 Primary Care Network 183790 Other 138789 Extended Access Provision 108122 Unmapped 27419
What is the number of context types?			Care Related Encounter 700481 Inconsistent Mapping 89494 Unmapped 27419

What is the number of national categories?			Inconsistent Mapping 89494 General Consultation Routine 89329 General Consultation Acute 84874 Planned Clinics 76429 Clinical Triage 74539 Planned Clinical Procedure 59631 Structured Medication Review 44467 Service provided by external 43095 Home Visit 41850 Unplanned Clinical Activity 40415 Patient contact during Care 28795 Unmapped 27419 Care Home Visit 26644 Social Prescribing Service 26492 Care Home Needs Assessment 23505 Non-contractual chargeable 20896 Walk-in 14179 Group Consultation and Group 5341
What is the number of appointment statuses?		Attended 232137 Unknown 201324 DNA 163360	
What is the date range of the provided data sets?	Min 2021-12-01 Max 2022-06-30	Min 2020-01-01 Max 2022-06-01	Min 2021-12-01 Max 2022-06-30
Which service setting reported the most appointments in North West London from 1 January to 1 June 2022?			The number of service settings for North West London General Practice 2104 Other 1318 Primary Care Network 1272 Extended Access Provision 1090 Unmapped 152
What is the number of appointments per month?	count_of_appo appointment_date appointment_date 2022 3 5 6 1 2 2021 12 2022 4		count_of_appointments appointment_date appointment_date 2021 11 30405070 10 30303834 2022 3 29595038 2021 9 28522501 2022 5 27495508 6 25828078 1 25635474 2 25355260 2021 12 25140776 2022 4 23913060 2021 8 23852171
What is the number of records per month?	appointment_date appointment_date 2021 12 19507 2022 1 19643 2 18974 3 21236 4 19078 5 20128 6 19227		appointment_date appointment_date 2021 8 69999 9 74922 10 74078 11 77652 12 72651 2022 1 71896 2 71769 3 82822 4 70012

			5 77425 6 74168
What was the actual utilisation of resources?		1013502.3	

Figure 3 – Most popular hasgtags

<table>
<tr>
<td>What are the top trending hashtags (#) on Twitter related to healthcare in the UK?</td>
<td>
<table>
<tr><th></th><th>Word</th><th>Count</th></tr>
<tr><td>0</td><td>#healthcare</td><td>716</td></tr>
<tr><td>1</td><td>#health</td><td>80</td></tr>
<tr><td>2</td><td>#medicine</td><td>41</td></tr>
<tr><td>3</td><td>#ai</td><td>40</td></tr>
<tr><td>4</td><td>#job</td><td>38</td></tr>
<tr><td>5</td><td>#medical</td><td>35</td></tr>
<tr><td>6</td><td>#strategy</td><td>30</td></tr>
<tr><td>7</td><td>#pharmaceutical</td><td>28</td></tr>
<tr><td>8</td><td>#digitalhealth</td><td>25</td></tr>
<tr><td>9</td><td>#pharma</td><td>25</td></tr>
</table>
</td>
</tr>
</table>

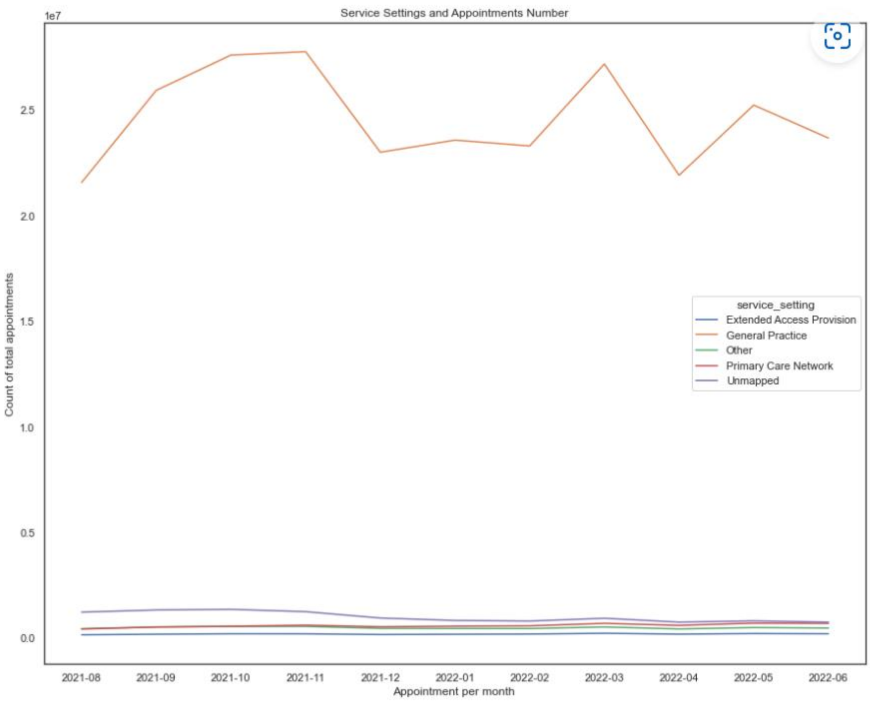
Figure 4 - NHS Network capacity per phashtagse and average utilization

Were there adequate staff and capacity in the networks?

	appointment_month	count_of_appointments	utilisation_percentage	utilisation_average
0	2021-08	23852171	8.1	795072.4
1	2021-09	28522501	9.6	950750.0
2	2021-10	30303834	10.2	1010127.8
3	2021-11	30405070	10.3	1013502.3
4	2021-12	25140776	8.5	838025.9
5	2022-01	25635474	8.7	854515.8
6	2022-02	25355260	8.6	845175.3
7	2022-03	29595038	10.0	986501.3
8	2022-04	23913060	8.1	797102.0
9	2022-05	27495508	9.3	916516.9
10	2022-06	25828078	8.7	860935.9

Figure 5 –

5.1. Service setting spread over appointment month



5.2. National categories spread over appointment month

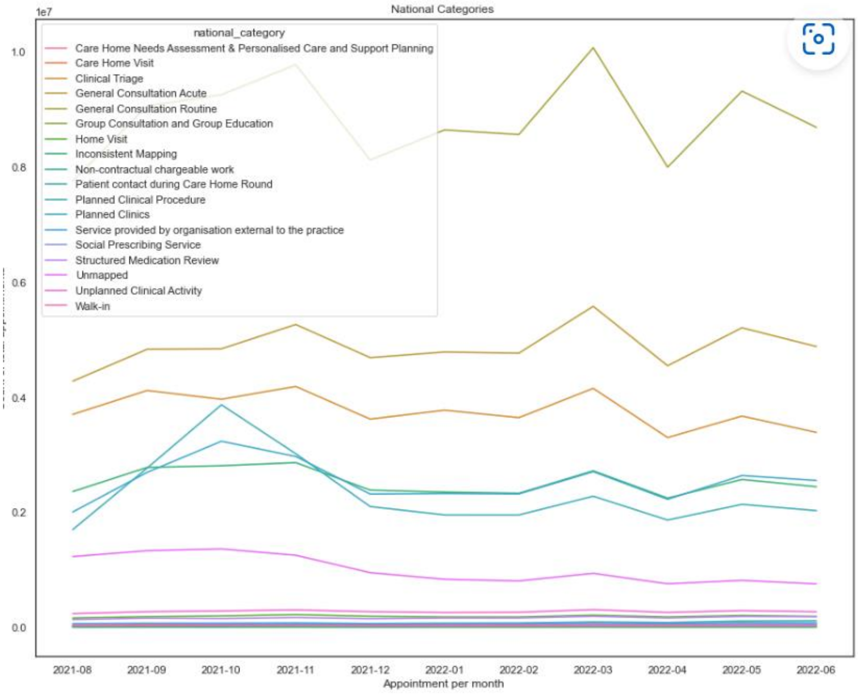
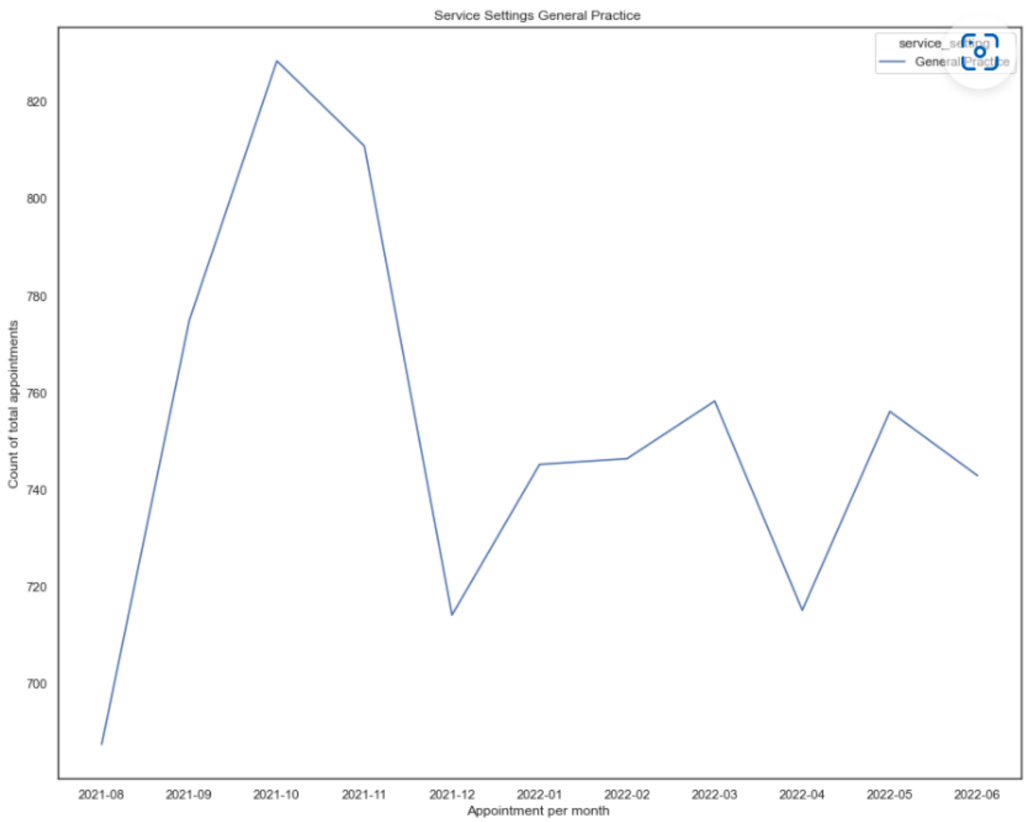


Figure 6 -

6.1. Service setting General Practice



6.2. Three service setting categories

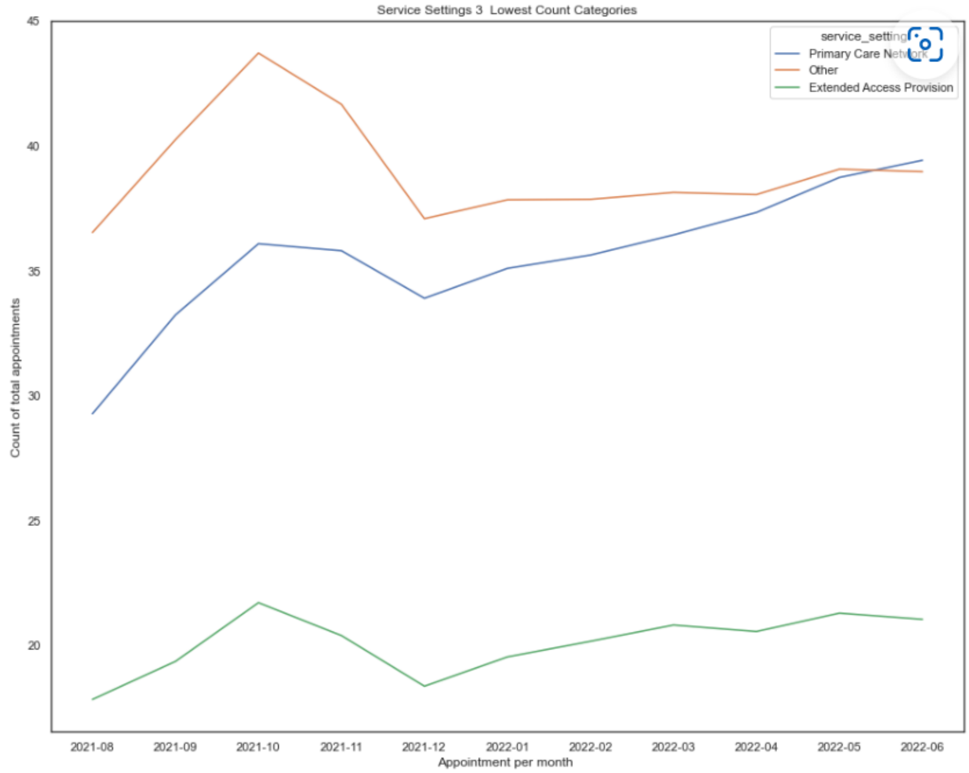
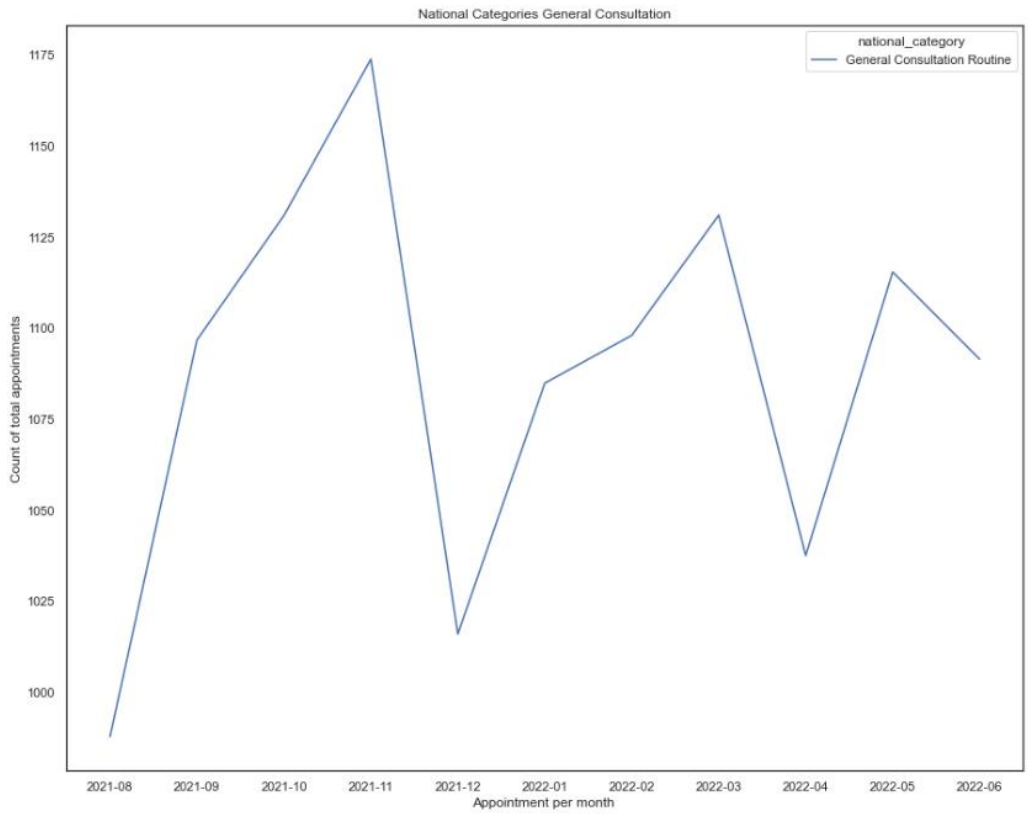


Figure 7 -

7.1. National Categories General Consultation



7.2. National categories six categories

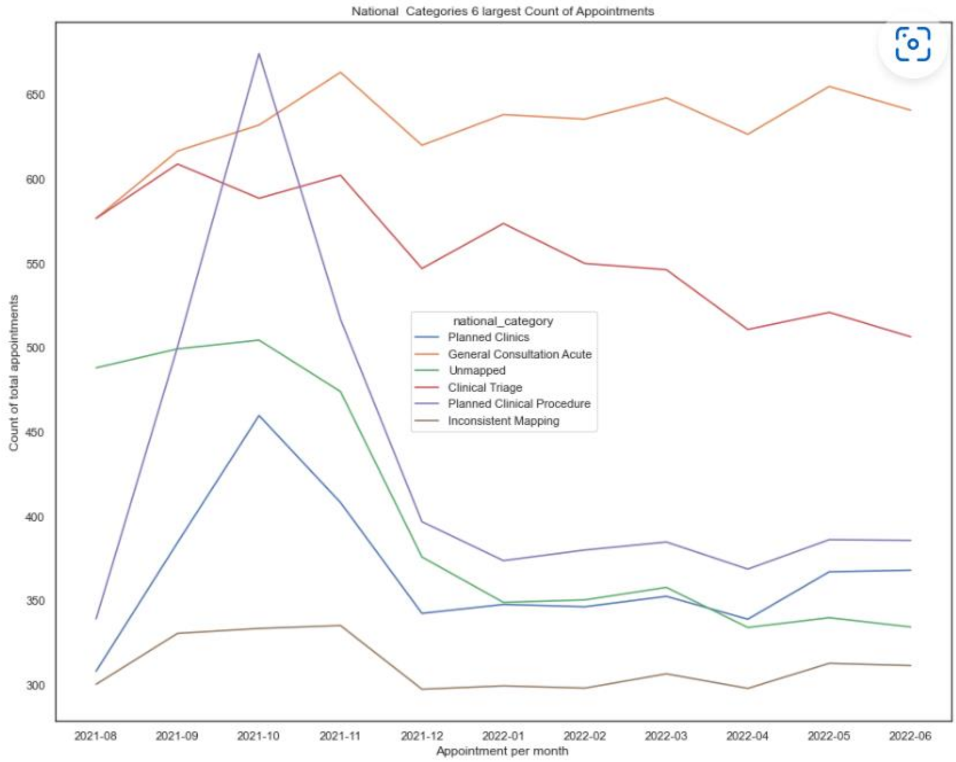
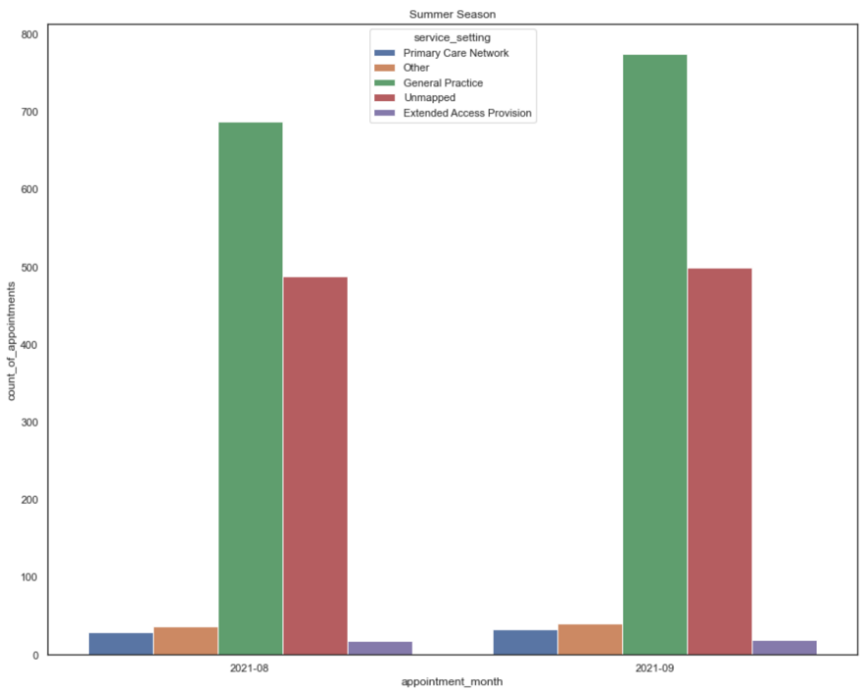
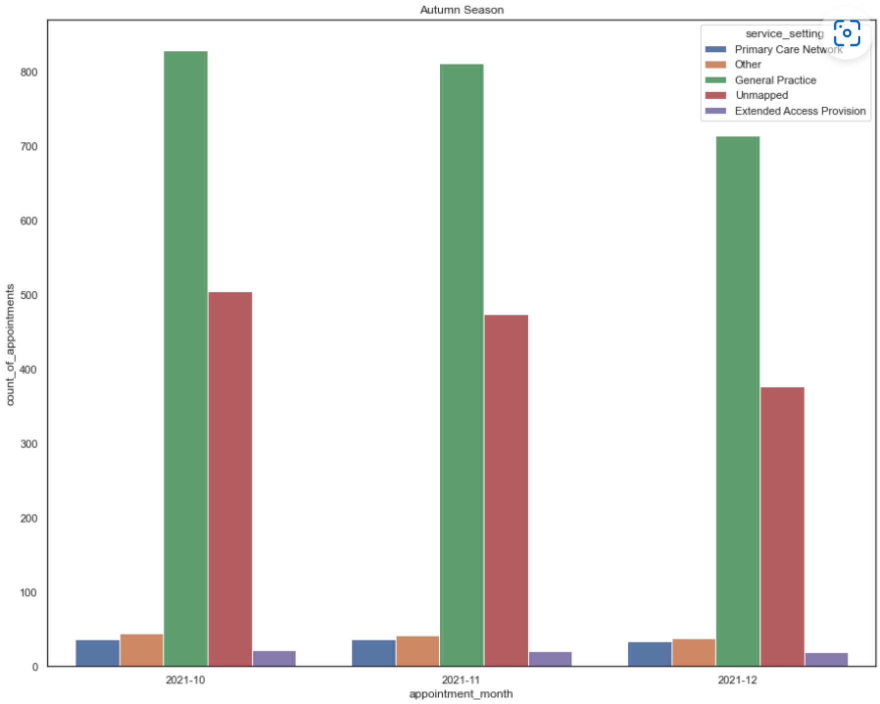


Figure 8

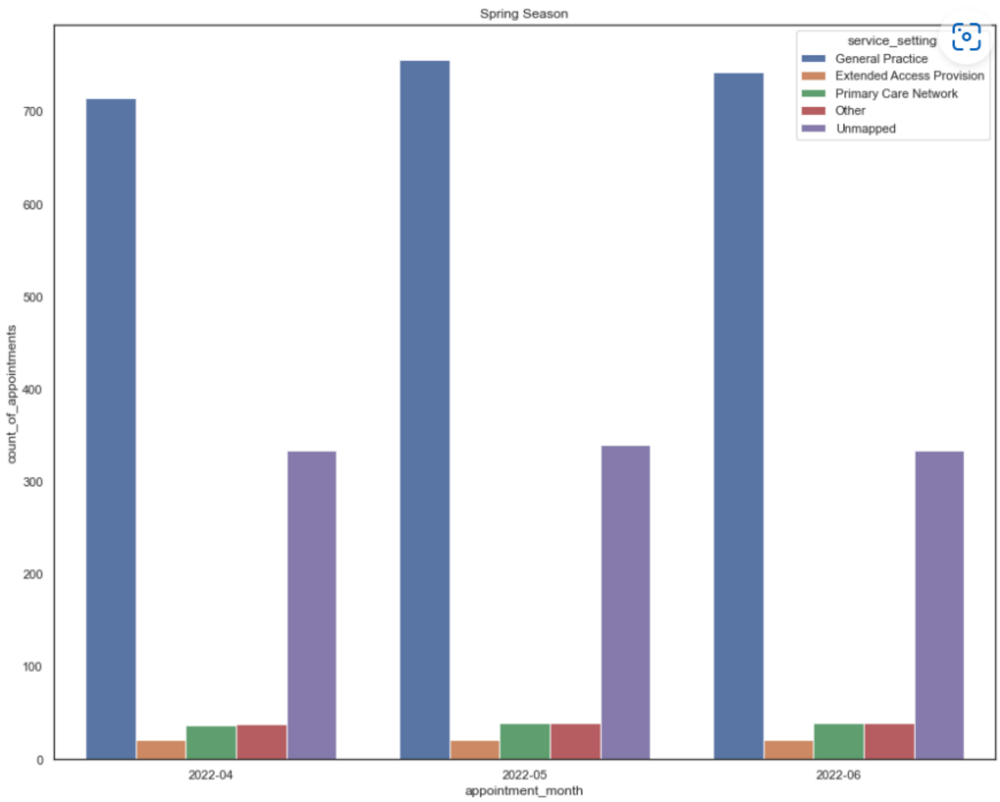
8.1. Summer Trend



8.2. Autumn Trend



8.3. Spring Trend



8.4. Winter Trend

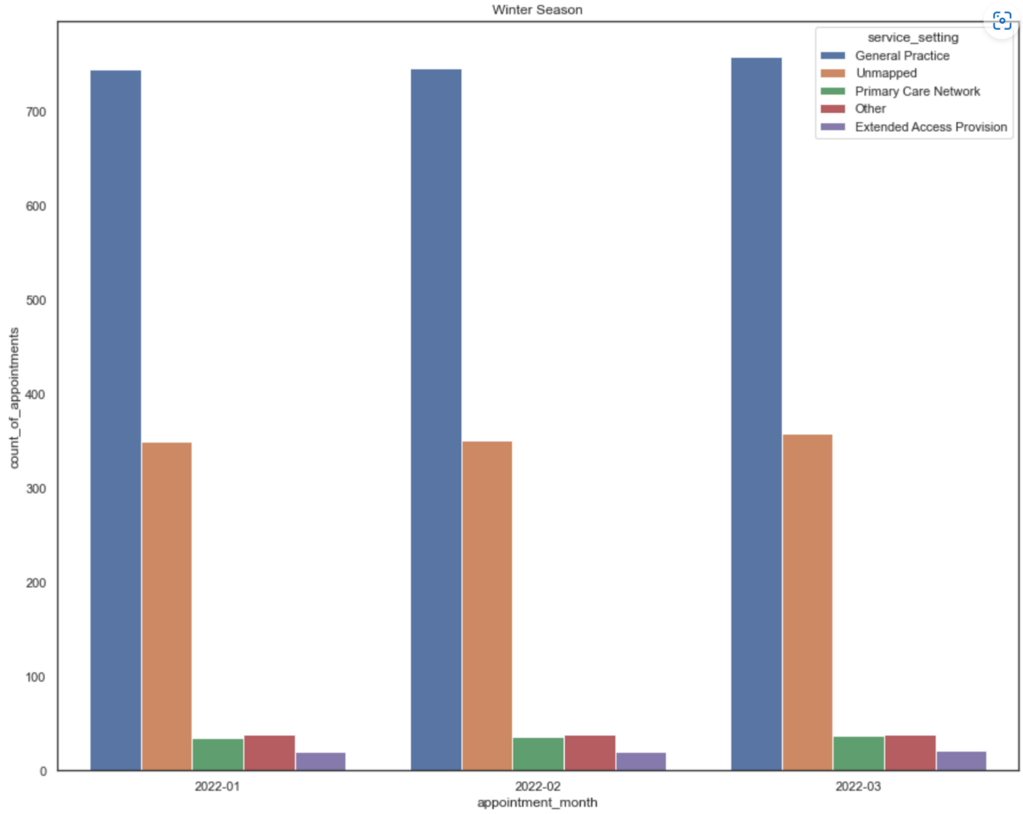


Figure 9 NHS most trending hashtags

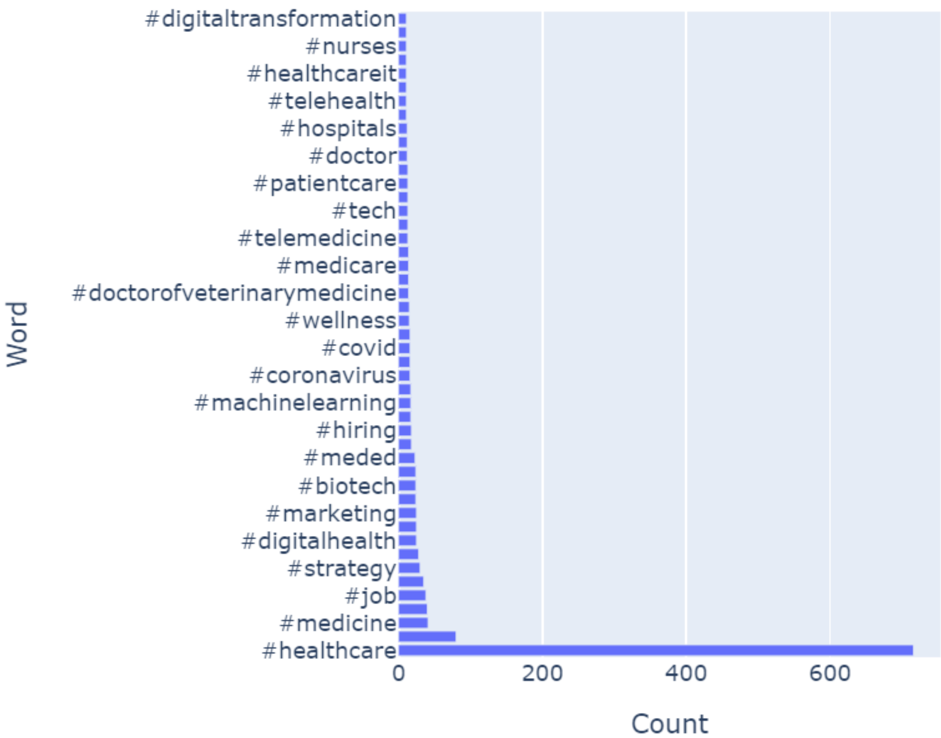


Figure 10 Bar plot of utilisation capacity

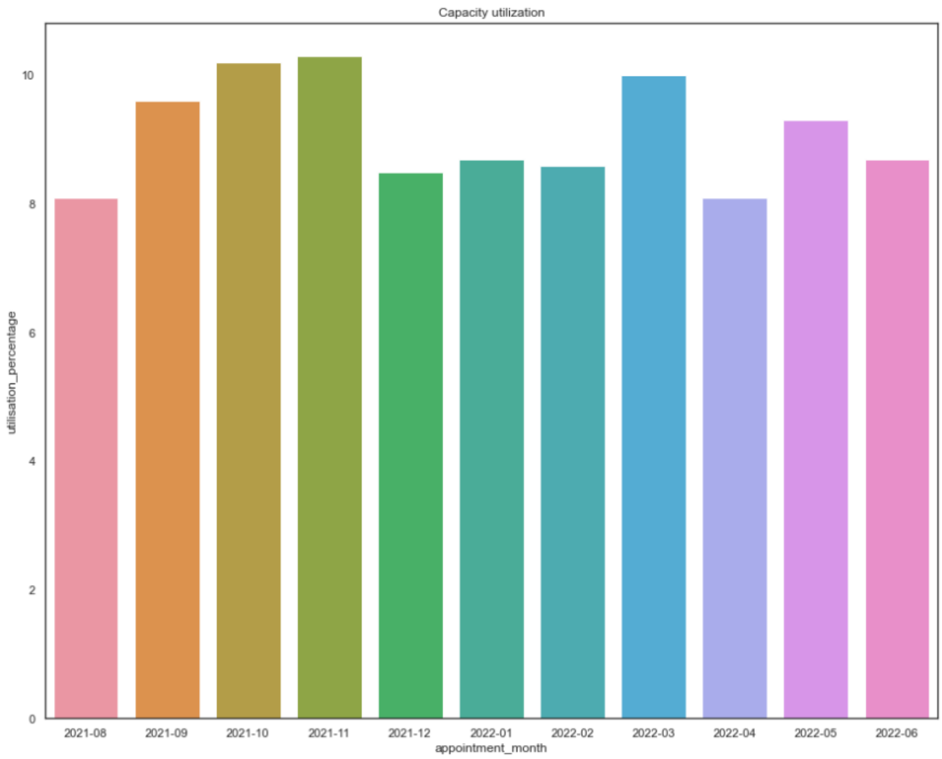


Figure 11 Healthcare professionals over time

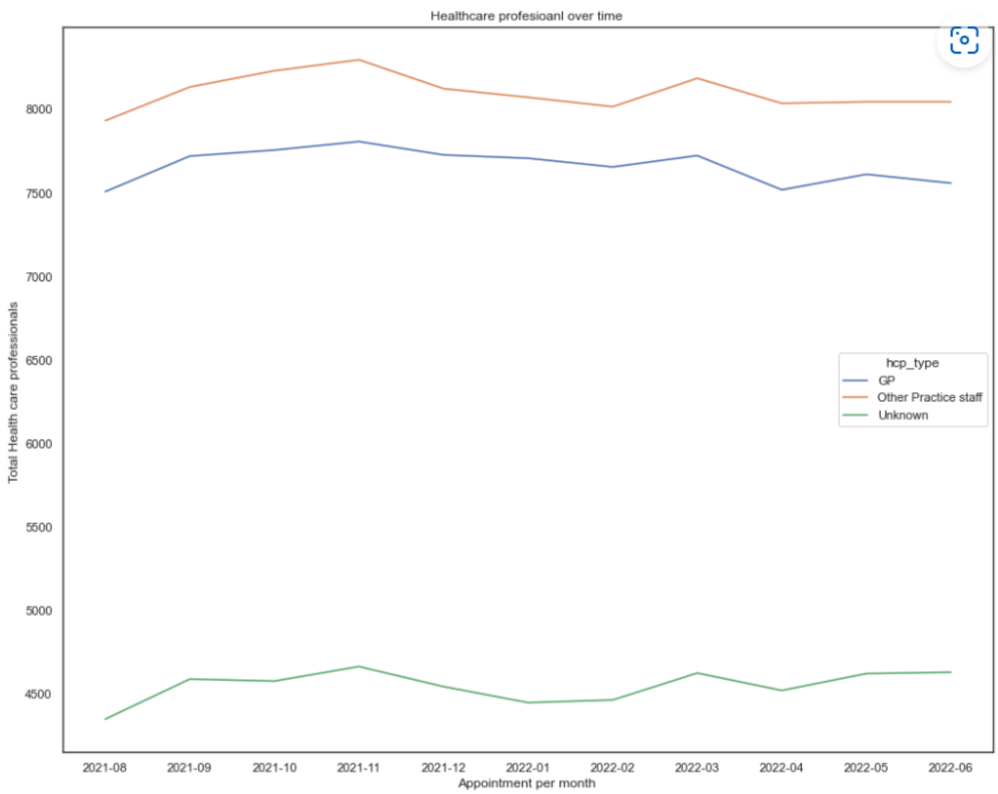


Figure 12 attended visits over time

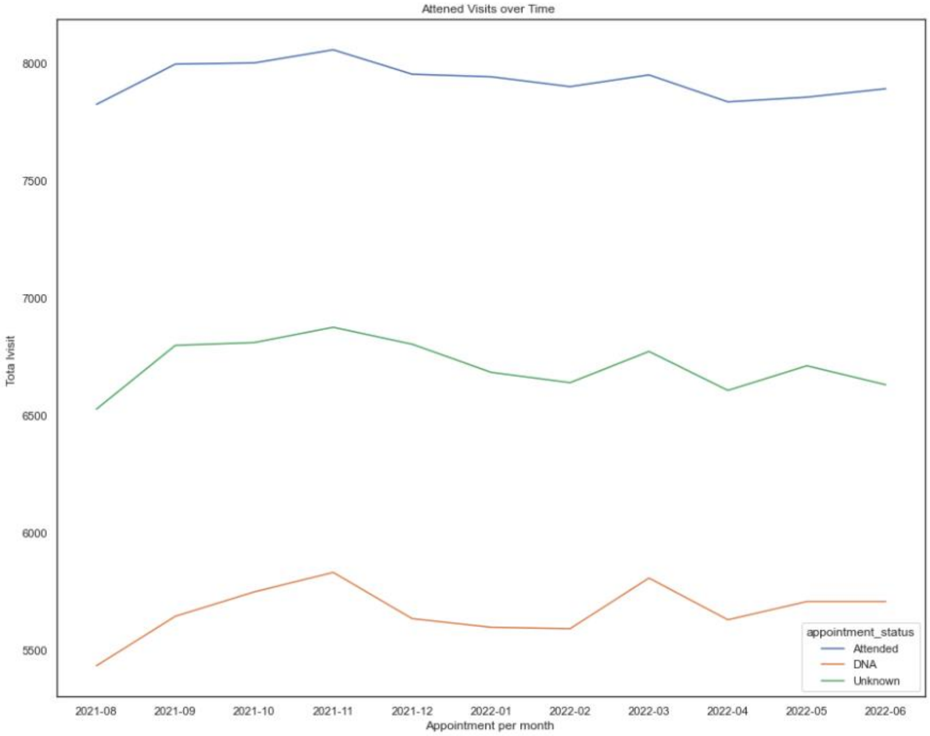


Figure 13 Appointment mode over time

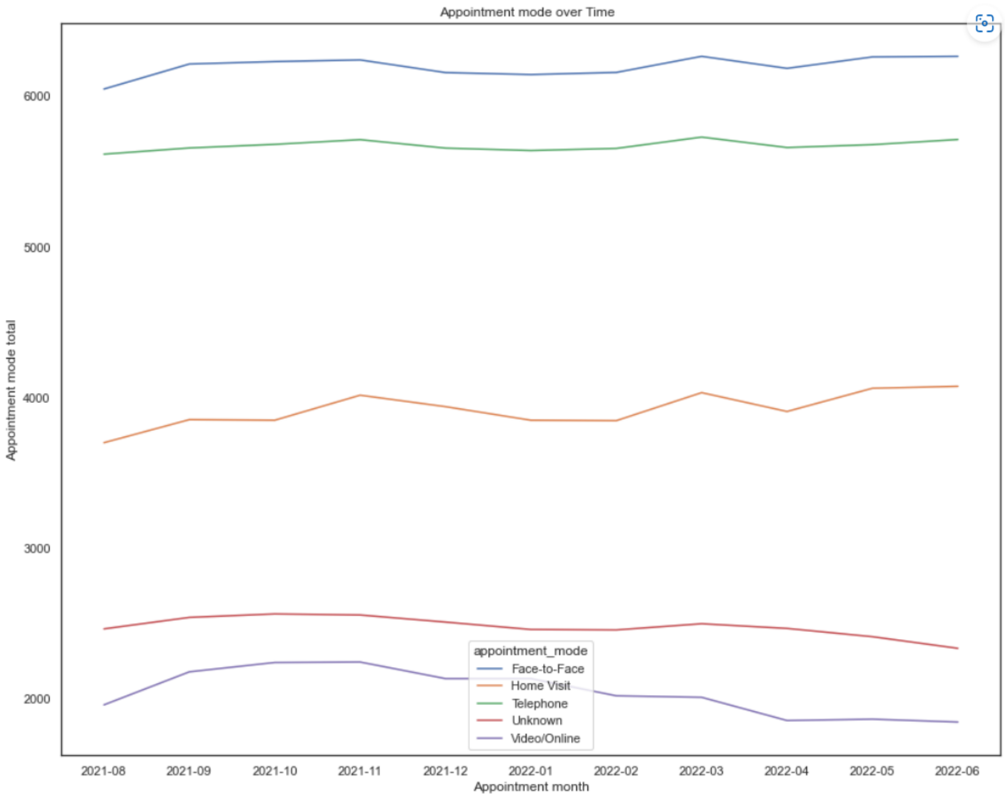


Figure 14 Spread of appointments over months

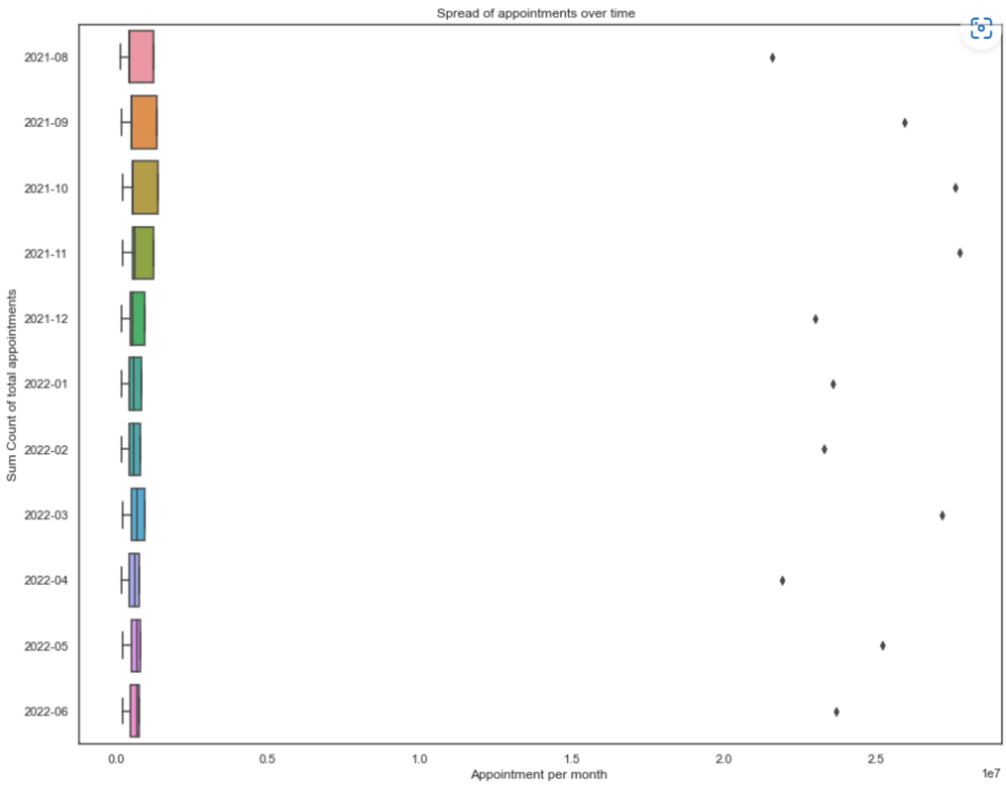
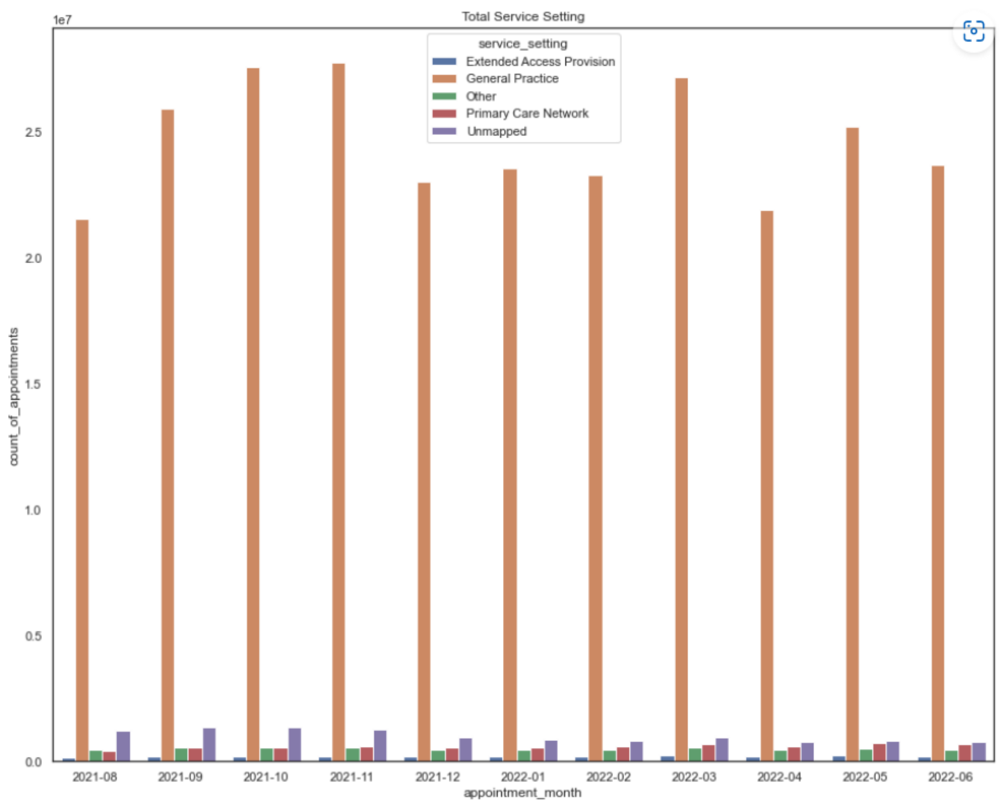


Figure 15 Service setting over time



15.1 Four categories of service settings