# Technical Report: Diagnostic Analysis Using Python on NHS Data.

## Background and Context

The NHS is looking to optimise its resources in terms of staff capacity and appointment attendance. Its focus is on analysing the cost of missed appointments to better allocate resources. The analysis focuses on answering the following questions:

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

The dataset provided included the following four datasets and their relative metadata:

- **actual_duration.csv** (imported as ad): provides the date, location, duration, and counts of appointments.
    - Location can be linked to examine the busiest locations.
    - Date helps visualise and analyse trends over time.
    - Count of appointments contextualises capacity.
- **appointments_regional.csv** (imported as ar): provides date, location, appointment status, appointment mode, healthcare professional, the time between booking an appointment, and the count of appointments.
    - Status helps explore attendance levels.
    - Mode looks at the type of appointment (ie. 'Face-to-Face, 'Telephone', etc).
    - Healthcare professionals look at staff capacity and use.
    - Time between booking and appointment helps investigate reasons for missed appointments.
- **national_categories.xlsx** (imported as nc): provides date, location, count of appointments, national categories, context type, and service setting.
    - National category frames the reason for the appointment (ie. routine, surgery (planned or unplanned), etc).
    - Service setting refers to a type of facility/location.
- **tweets.csv** (imported as tweet): provides NHS-related tweets to give a sentiment review over the topic.

## Analytical Approach

The first steps in the analysis are to review, confirm, and explore the data. There are no missing values but the exploration shows a large number of 'unmapped' or 'inconsistent mapping' values which hinder the quality of the set and the accuracy of the analysis.

Nonetheless, the initial exploration showed that there are 5 service settings, 3 context types, 18 national categories, and 3 types of appointment statuses:

```
And the service settings are:
General Practice              359274
Primary Care Network          183790
Other                         138789
Extended Access Provision     108122
Unmapped                       27419

The context types are:
Care Related Encounter    700481
Inconsistent Mapping       89494
Unmapped                   27419


The national categories are:
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 organisation external to the practice             43095
Home Visit                                                            41850
Unplanned Clinical Activity                                           40415
Patient contact during Care Home Round                                28795
Unmapped                                                              27419
Care Home Visit                                                       26644
Social Prescribing Service                                            26492
Care Home Needs Assessment & Personalised Care and Support Planning   23505
Non-contractual chargeable work                                       20896
Walk-in                                                               14179
Group Consultation and Group Education                                 5341


The appointment statuses are:
Attended    232137
Unknown     201324
DNA         163360
```

Furthermore, the datasets don't have all the same data collection start.

```
# Determine the minimum and maximum dates in the ad DataFrame.
# Use appropriate docstrings.
min_ad = ad['appointment_date'].min()
max_ad = ad['appointment_date'].max()

print("The appointments data was collected from %s to %s " % (min_ad, max_ad))
```

```
The appointments data was collected from 2021-12-01 00:00:00 to 2022-06-30 00:00:00
```

```
# Determine the minimum and maximum dates in the nc DataFrame.
# Use appropriate docstrings.
min_nc = nc['appointment_date'].min()
max_nc = nc['appointment_date'].max()

print("The appointments data was collected from %s to %s" % (min_nc, max_nc))
```

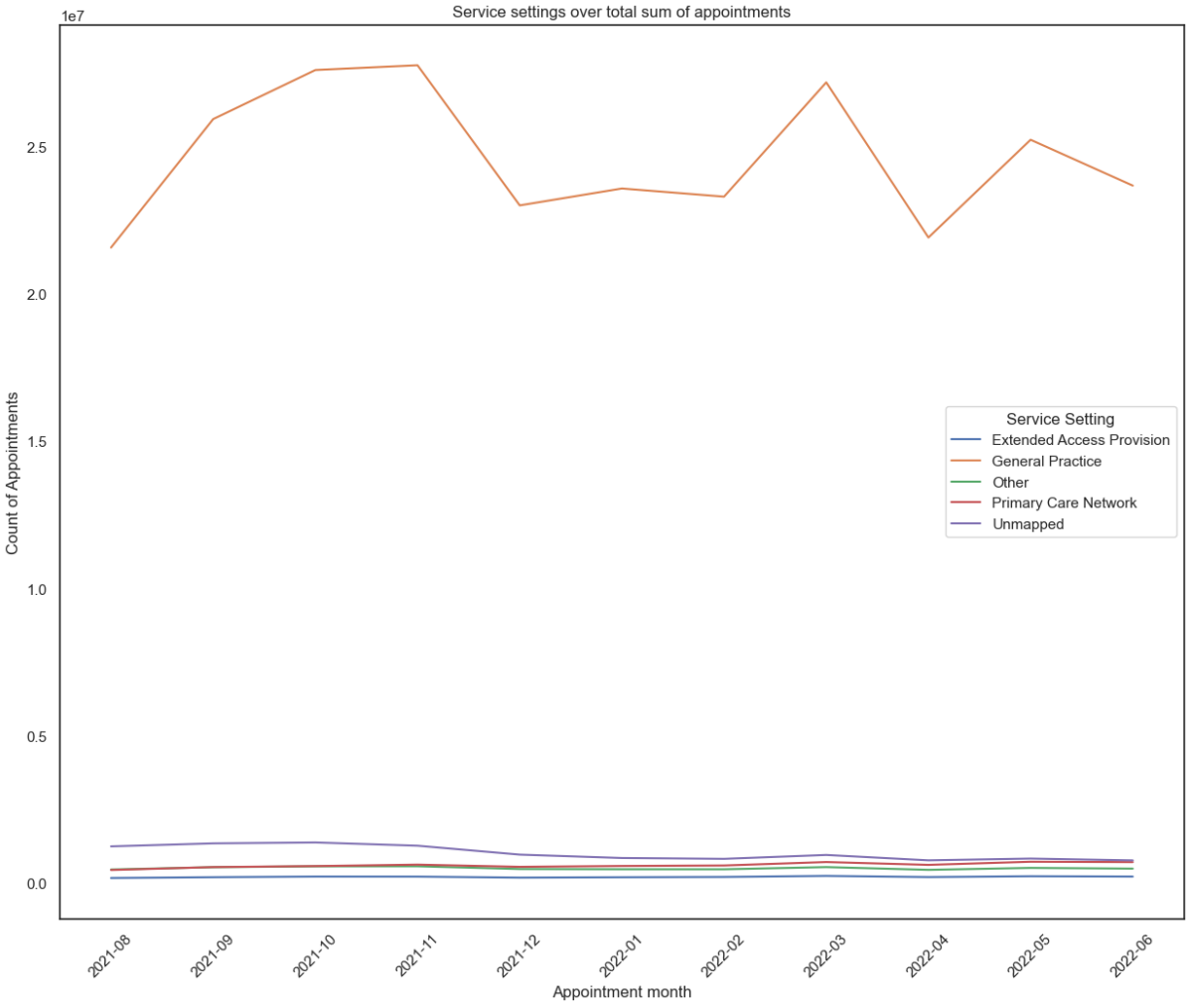
```
The appointments data was collected from 2021-08-01 00:00:00 to 2022-06-30 00:00:00
```

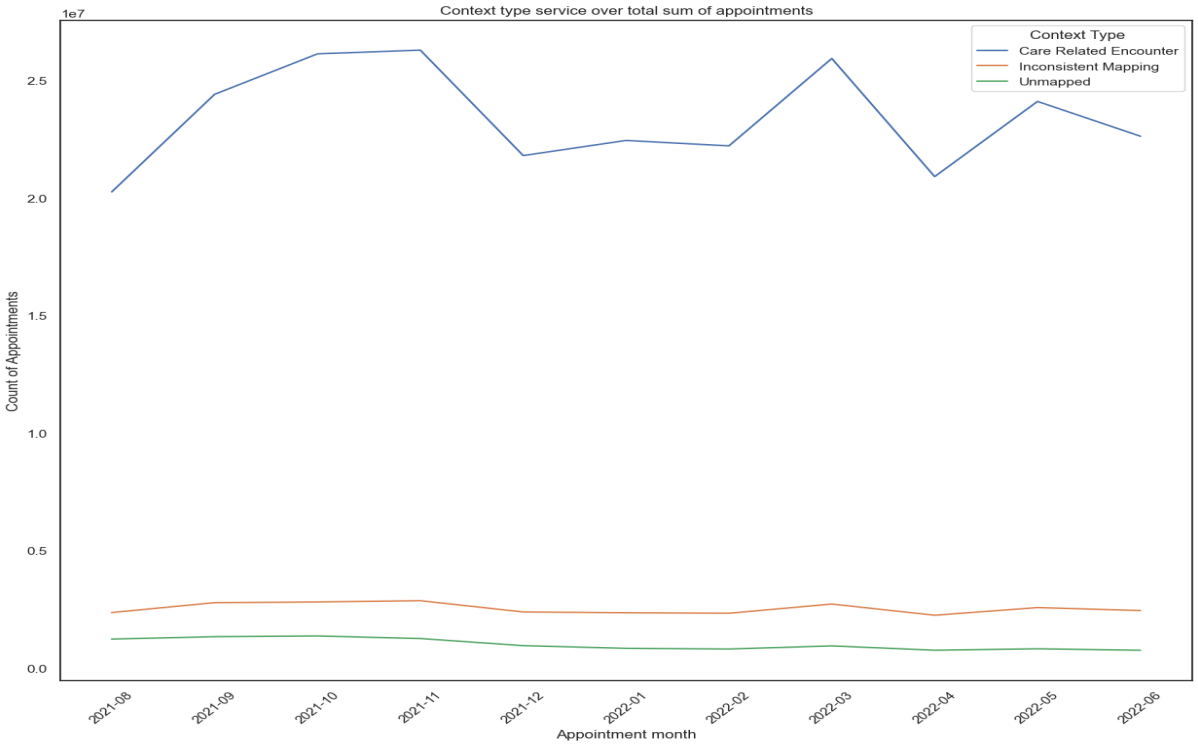
The min and max for the 'national category' and 'appointment duration' datasets show that the ranges of data collections are not the same length. Data from the 'national category' dataset start in August 2021, while the 'appointment duration' dataset starts in December 2021. The data range ends in June 2022 for both sets.

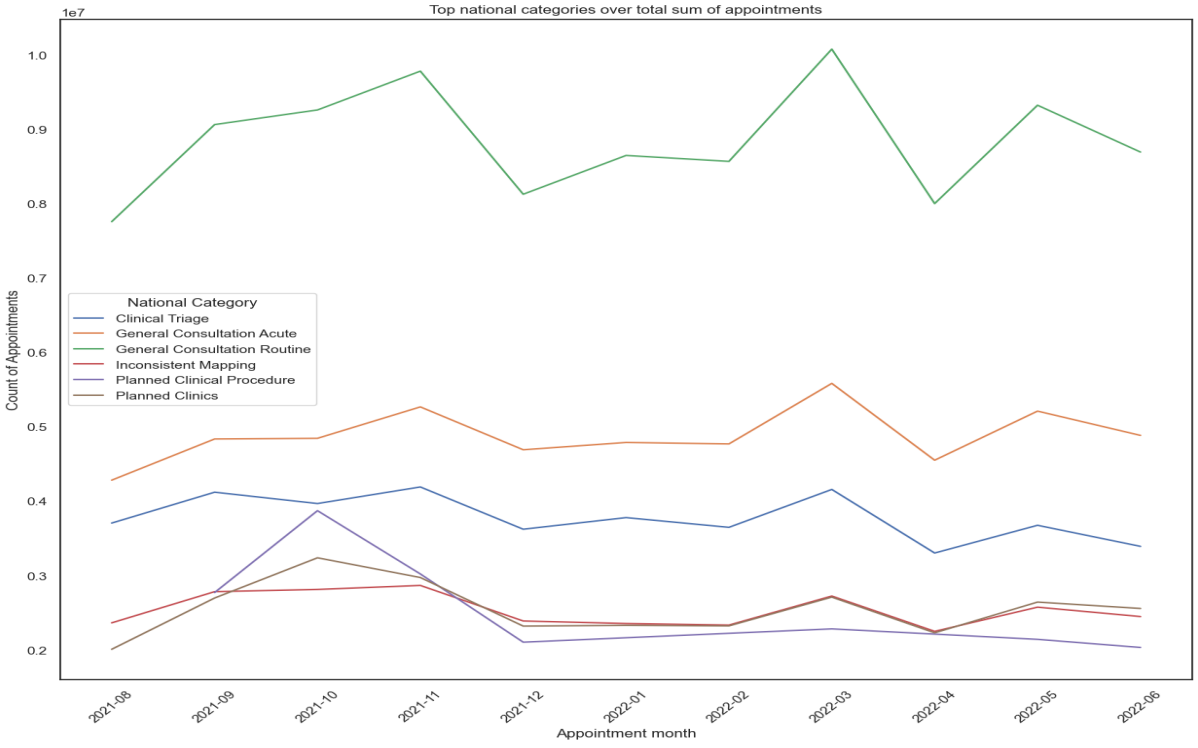
After confirming the datasets, the analysis focused on breaking down the initial questions into smaller parts:

- Identifying the busiest month both on count of appointments and number of records. Considering that the ad dataset starts from December 2021, more attention was given to the nc one.
    - It shows November 2021 as the busiest month in terms of counts while in terms of records, it is March 2022, with August 2021 as the least busy month in both cases.

Keeping that in mind, the analysis progressed to plotting the trends in service settings, context types, and national categories over the dataset timeframe.

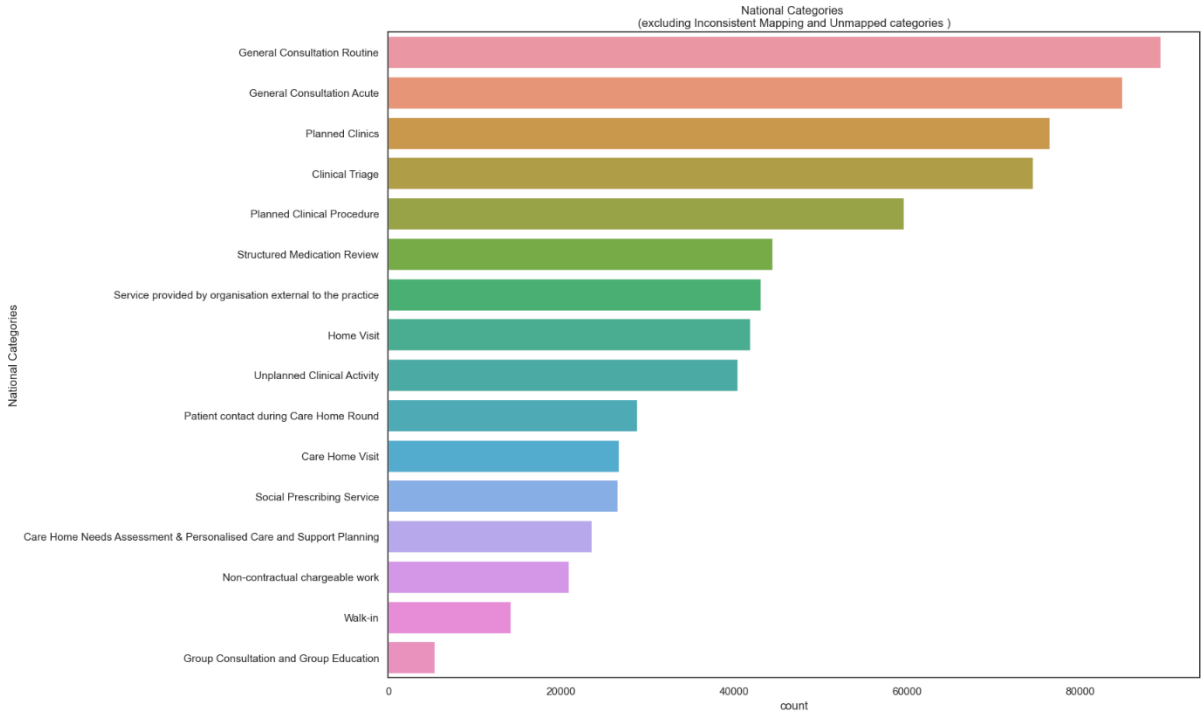


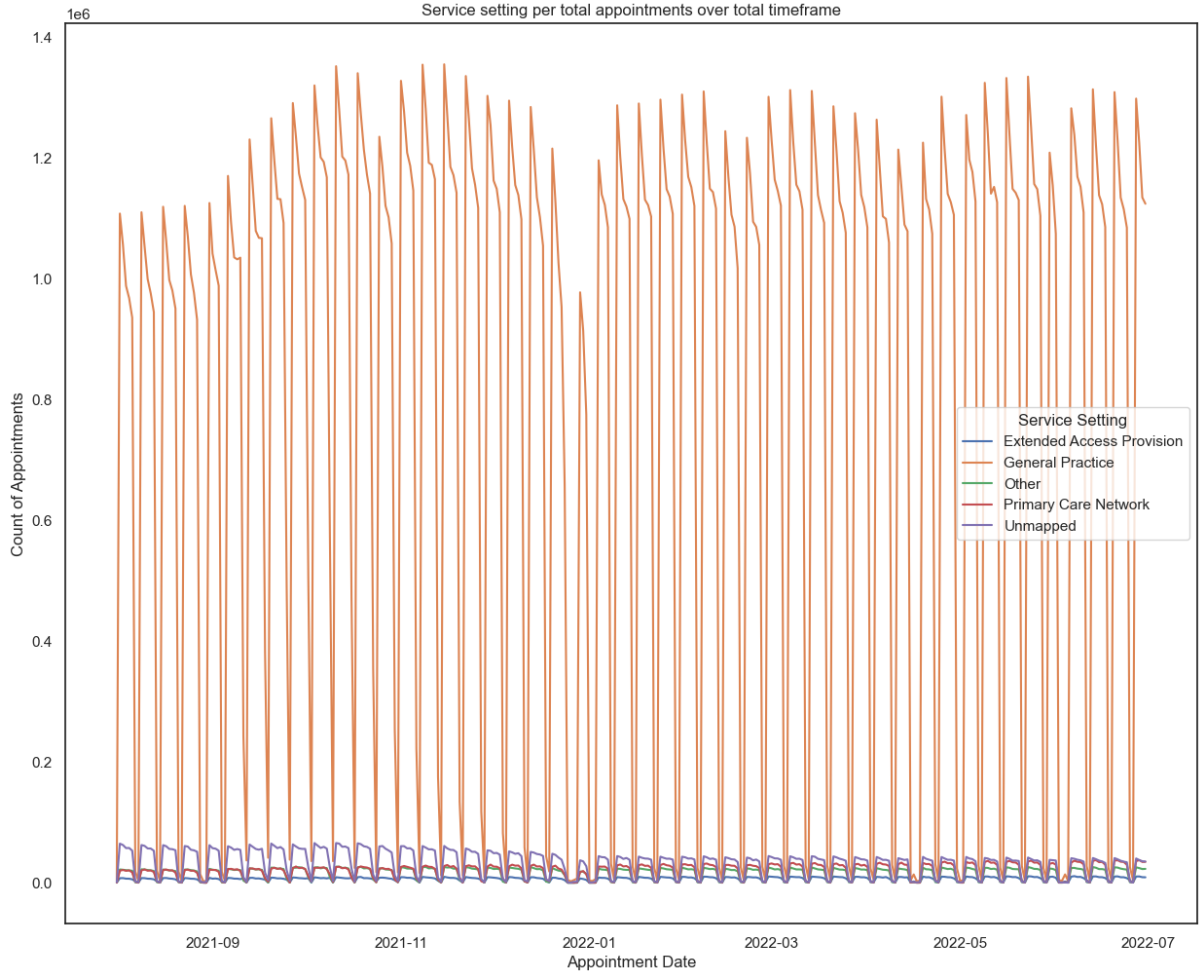



The three plots confirm the busiest periods already calculated and show:

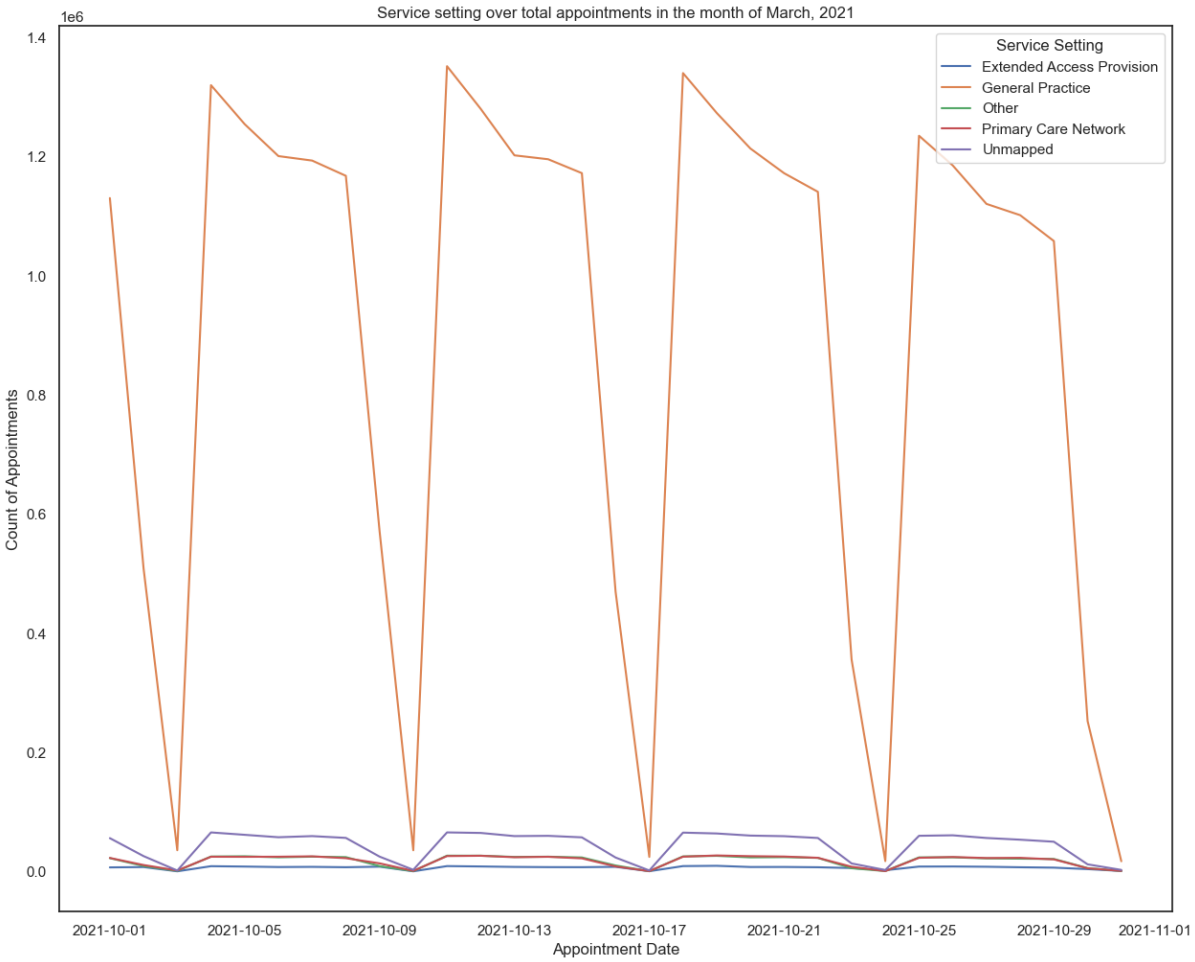
- The GP Practices are far above the most used service setting.
- Care Related Encounters are the most frequent type of context.
- General Routine Consultations are the most frequent type of appointment, followed by General Consultation Acute.


National Categories
(excluding Inconsistent Mapping and Unmapped categories )
General Consultation Routine
General Consultation Acute
Planned Clinics
Clinical Triage
Planned Clinical Procedure
Structured Medication Review
Service provided by organisation external to the practice
Home Visit
Unplanned Clinical Activity
Patient contact during Care Home Round
Care Home Visit
Social Prescribing Service
Care Home Needs Assessment & Personalised Care and Support Planning
Non-contractual chargeable work
Walk-in
Group Consultation and Group Education
National Categories
0
20000
40000
60000
80000
count

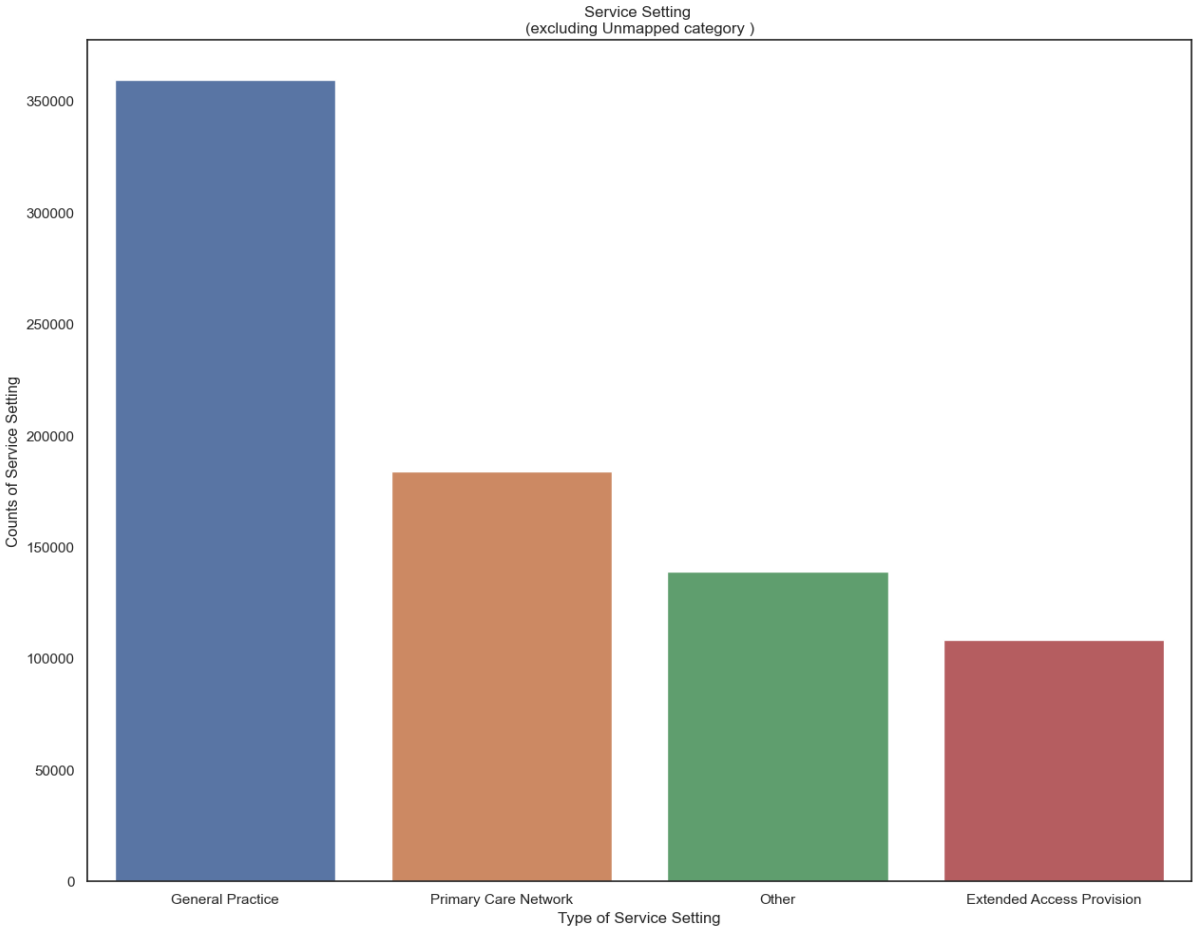

Service setting per total appointments over total timeframe
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count of Appointments
Service Setting
Extended Access Provision
General Practice
Other
Primary Care Network
Unmapped
2021-09
2021-11
2022-01
2022-03
2022-05
2022-07
Appointment Date

While the bar chart confirms in a bar plot the national categories, the overall trends in service settings show a more complete view of the count of appointments for the time span and highlights the busiest months vs the least busy ones, such as Jan/Feb 2022.

Looking more closely, we can see the trends of appointments peaking at the beginning of the week and declining as the days go by, as shown by the line plot for the service setting in the month of March 2022.




By excluding the ‘unmapped’ category we can have a clearer overview of the service settings, with GP still the highest setting, but with ‘Primary Care Network’ and ‘Other’ as an important part.

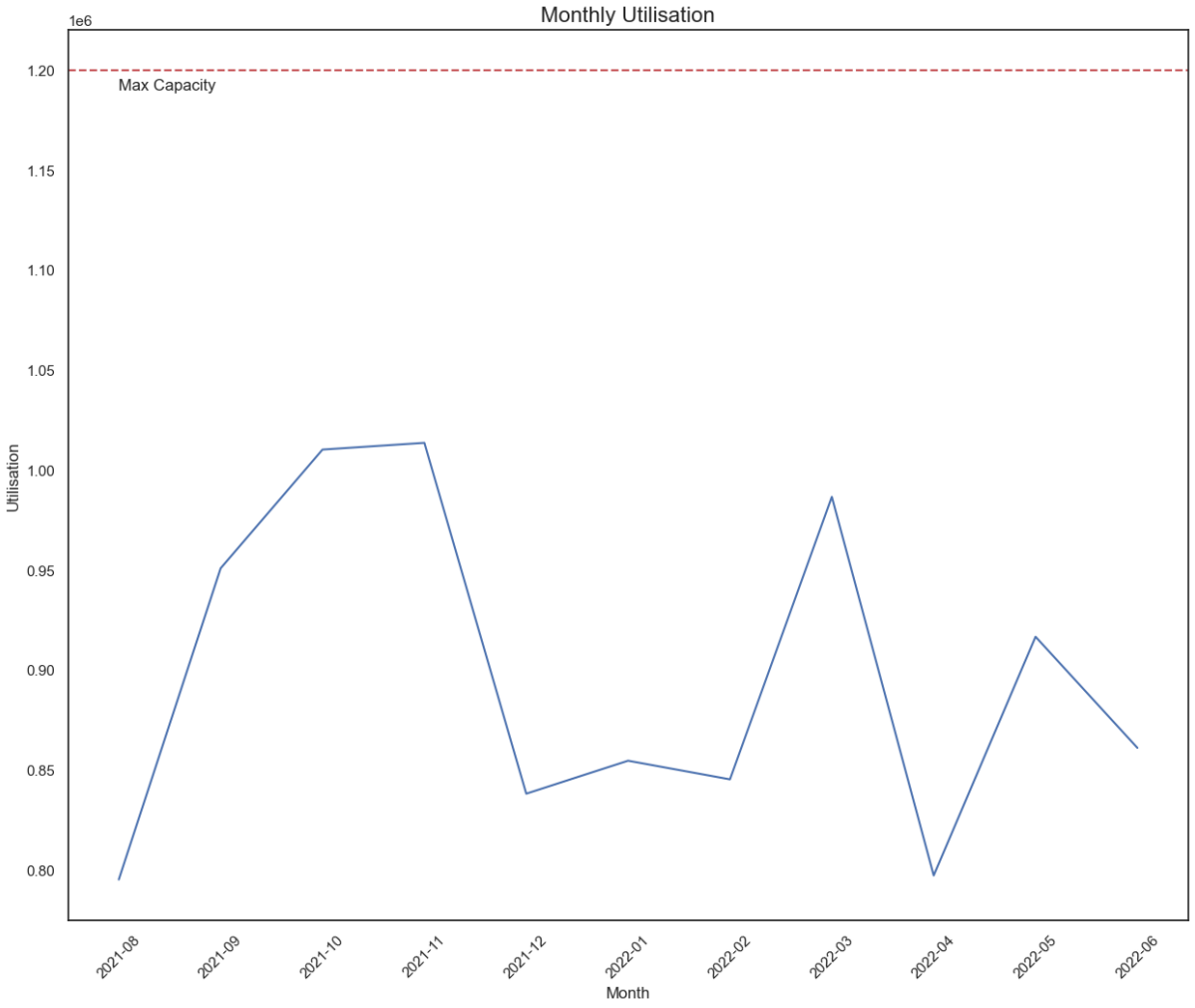



While the tweets data set has been explored, it did not generate particularly relevant insights for the analysis, but it does show interest and discussion around the healthcare topic.

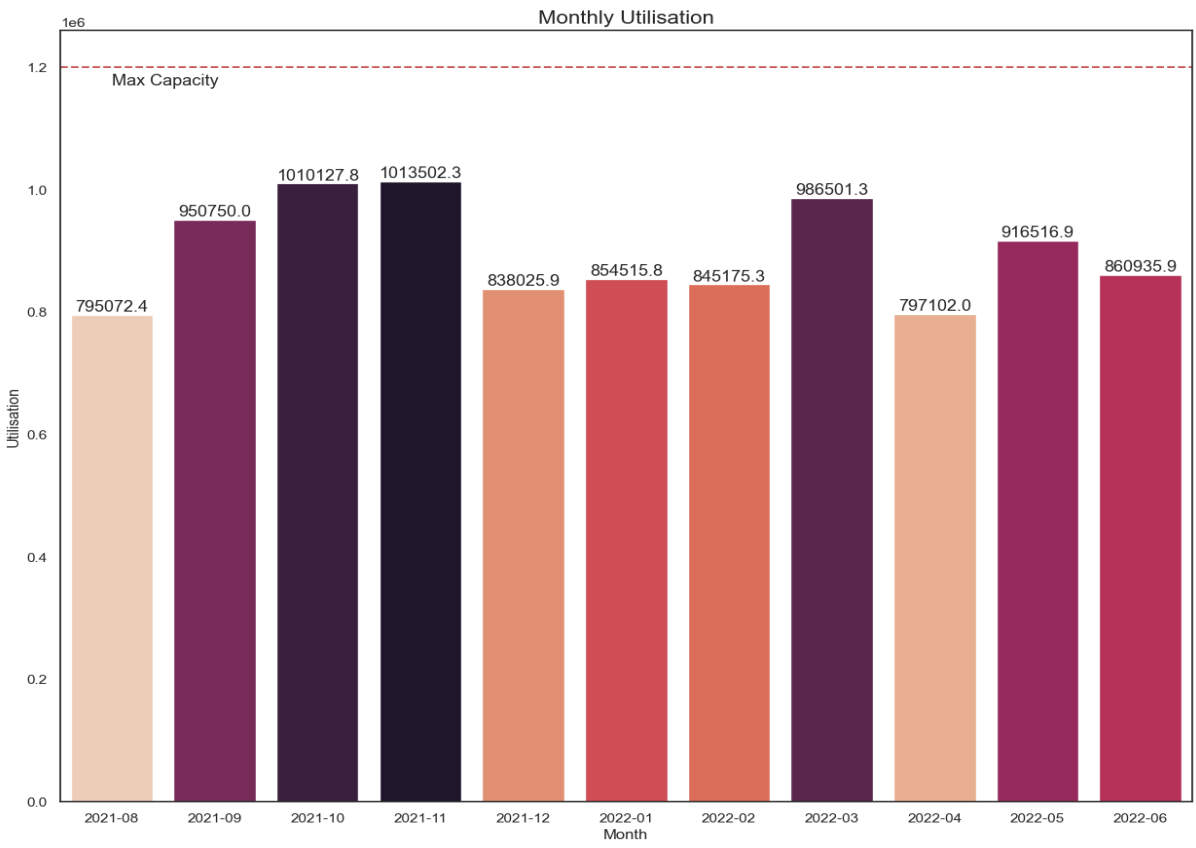
## Trends and Insights

The analytical exploration provided the context to individuate noticeable trends and the framework to generate insights. As such, the analysis aimed to answer issues of staff capacity and utilisation.  By plotting the volume of appointments and the max capacity of the NHS network, it was possible to see that the capacity levels were below the max, even in the busiest months. Therefore, the issue doesn't lie in the capacity of the NHS network but in the distribution of resources and the staff available, as well as the booking times and availabilities.

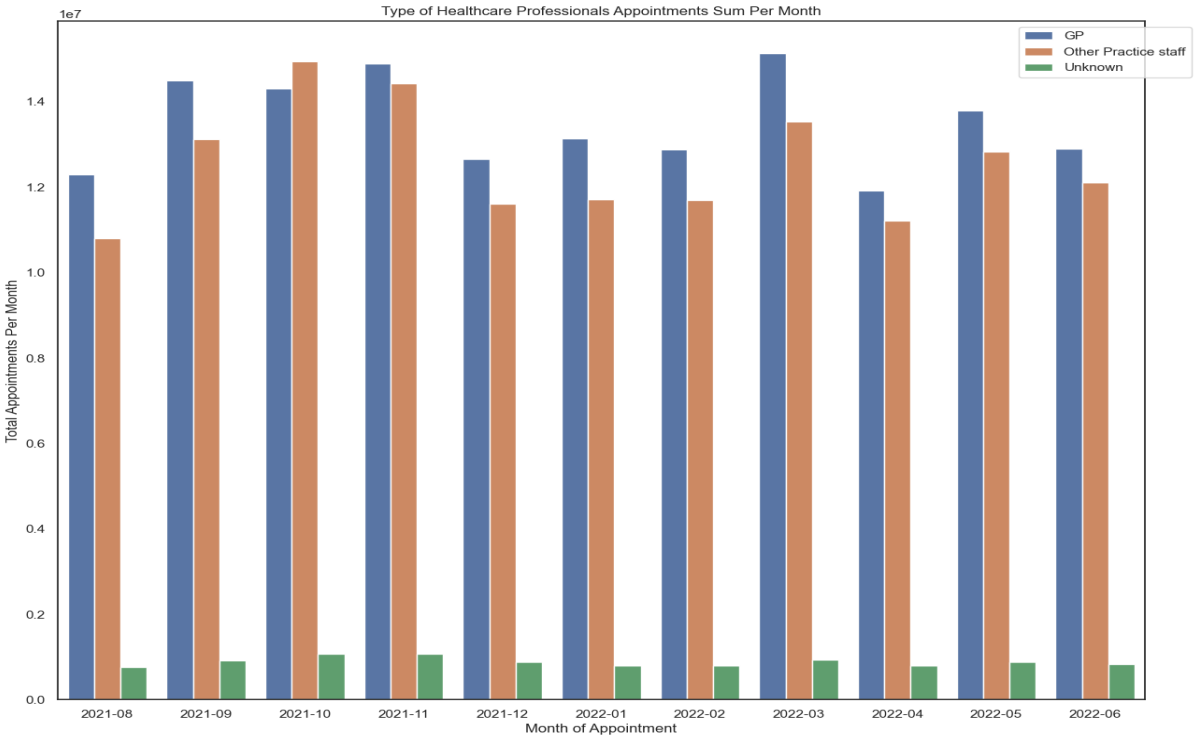
```
# Plot monthly capacity utilisation.
max_capacity = 1200000
labels = ar_app['utilisation'].tolist()

# Create a lineplot.
sns.lineplot(data=ar_app, x=ar_app['appointment_month'], y=ar_app['utilisation'])
plt.title("Monthly Utilisation",fontsize=16)
plt.xlabel("Month")
plt.ylabel("Utilisation")
plt.xticks(rotation=45)
plt.axhline(y=max_capacity, color='r', linestyle='--')
plt.text('2021-08',1190000, "Max Capacity")
plt.show
```




```
# Barplot with values and shade
labels = ar_app['utilisation'].tolist()  # Labels on bars
pal = sns.color_palette("rocket", len(ar_app)) # Palette for shaded bars
rank = ar_app['utilisation'].argsort().argsort()
chart = sns.barplot(data=ar_app,y=ar_app['utilisation'], x=ar_app['appointment_month'], palette=np.array(pal[::-1])[rank])
plt.title("Monthly Utilisation",fontsize=16)
plt.xlabel("Month")
plt.ylabel("Utilisation")
plt.text(0.0001,1170000, "Max Capacity", fontsize=13)
plt.axhline(y=max_capacity, color='r', linestyle='--')
chart.bar_label(chart.containers[0], labels=labels, label_type='edge', size=13)  # Setting the values on the bars
```




Furthermore, by plotting the different healthcare professionals we see how to better utilise the staff and provide a more complete cover over the NHS network.

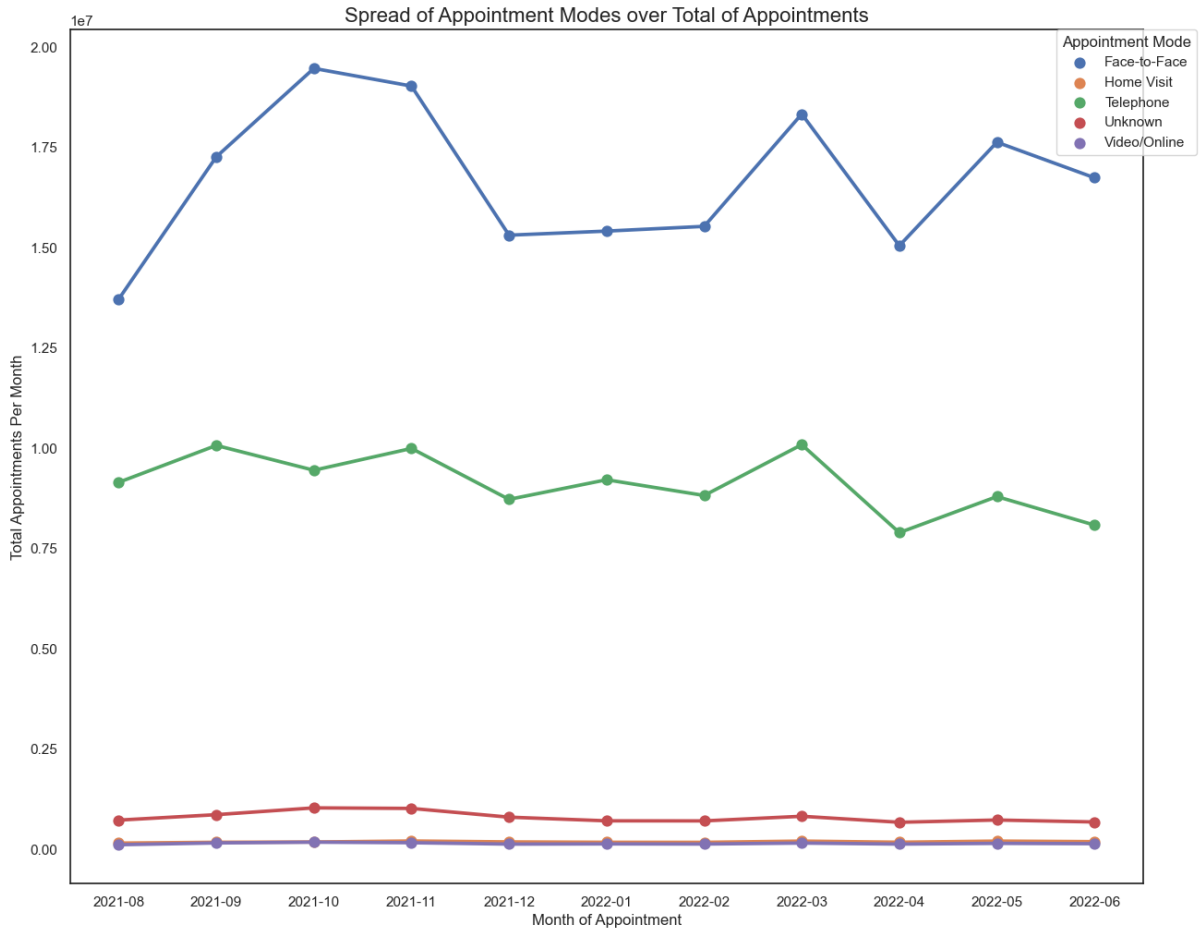


```
# Barplot
sns.barplot(x='appointment_month', y='count_of_appointments', hue='hcp_type', data=ar_hcp)
plt.title("Type of Healthcare Professionals Appointments Sum Per Month")
plt.xlabel('Month of Appointment')
plt.ylabel('Total Appointments Per Month')
plt.legend(bbox_to_anchor=(1.05, 1), loc='best')
plt.show
```

The bar chart shows that in the month of October 2021 (one of the busiest months) 'other practice staff' surpassed the GP in the total of appointments given. It shows that support given to the GP practices can be varied and can support the overall network.

By analysing further the appointment modes, we can better look at how to optimise utilisation.

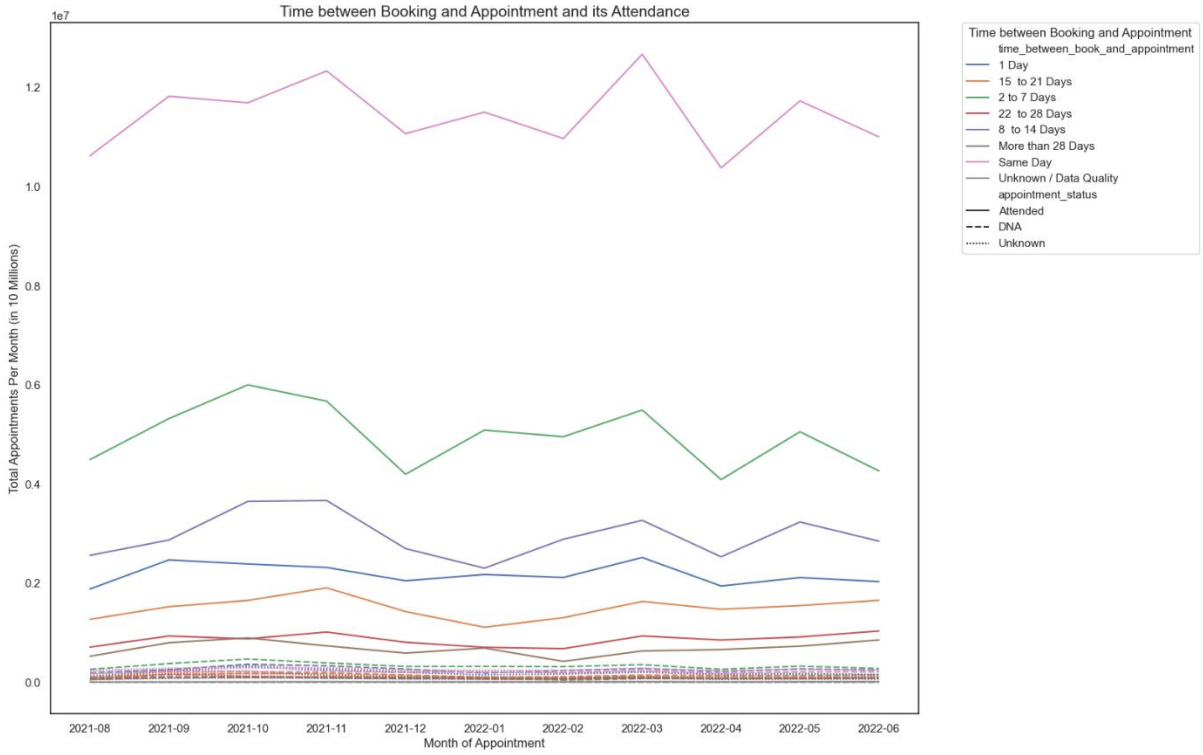
```
sns.pointplot(x='appointment_month', y='count_of_appointments', hue='appointment_mode',data=ar_mode)
plt.title("Spread of Appointment Modes over Total of Appointments", fontsize=16)
plt.xlabel('Month of Appointment')
plt.ylabel('Total Appointments Per Month')
plt.legend(title="Appointment Mode", bbox_to_anchor=(1.05, 1), loc='best', borderaxespad=0)
```



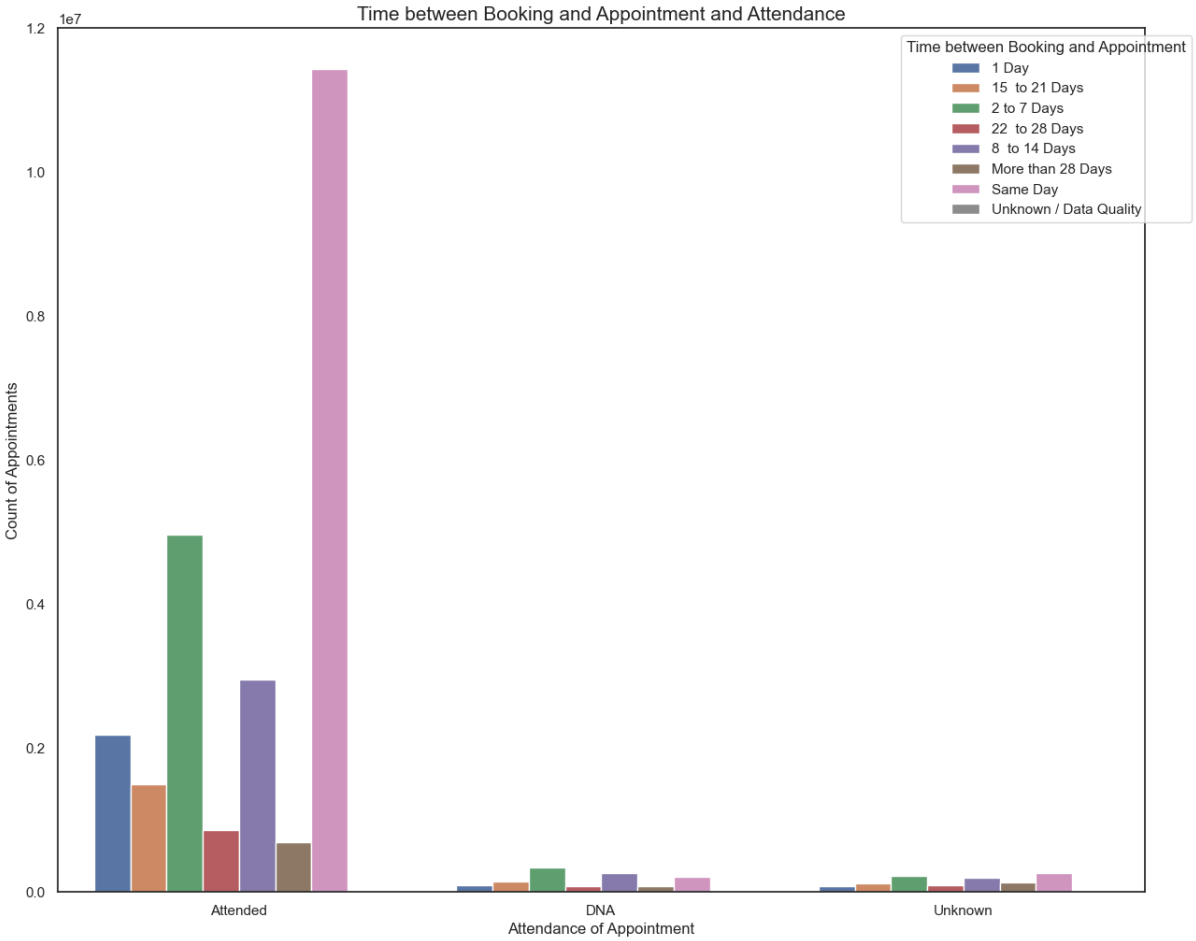

The line plot with points shows that in October 2021 as 'Face-to-Face' appointments rose, 'Telephone' ones decreased, only to increase again the following month, as November 2021 was the busiest time. The second consideration is that in the least busy period (Jan/Feb 2022) 'Telephone' appointments slightly rose. This shows that 'Telephone' appointments provide good support when capacity is low and high and can be an element to further enhance.

Looking into the length between booking and appointments, most appointments are booked on the day, followed by ‘2 to 7 days’ and ‘8 to 14 days’. When looking at the missed appointments it shows that ‘2 to 7 days' is the category most likely to be missed. Considering they make up less than half the total number of appointments it is a significant percentage.

```
# Correlation between booking time and attendance?
sns.lineplot(x='appointment_month', y='count_of_appointments', hue='time_between_book_and_appointment',
             style='appointment_status',data=ar_time)
plt.title("Time between Booking and Appointment and its Attendance", fontsize=15)
plt.xlabel('Month of Appointment')
plt.ylabel('Total Appointments Per Month')
plt.legend(title="Time between Booking and Appointment", bbox_to_anchor=(1.05, 1), loc='best', borderaxespad=0)
plt.show
```



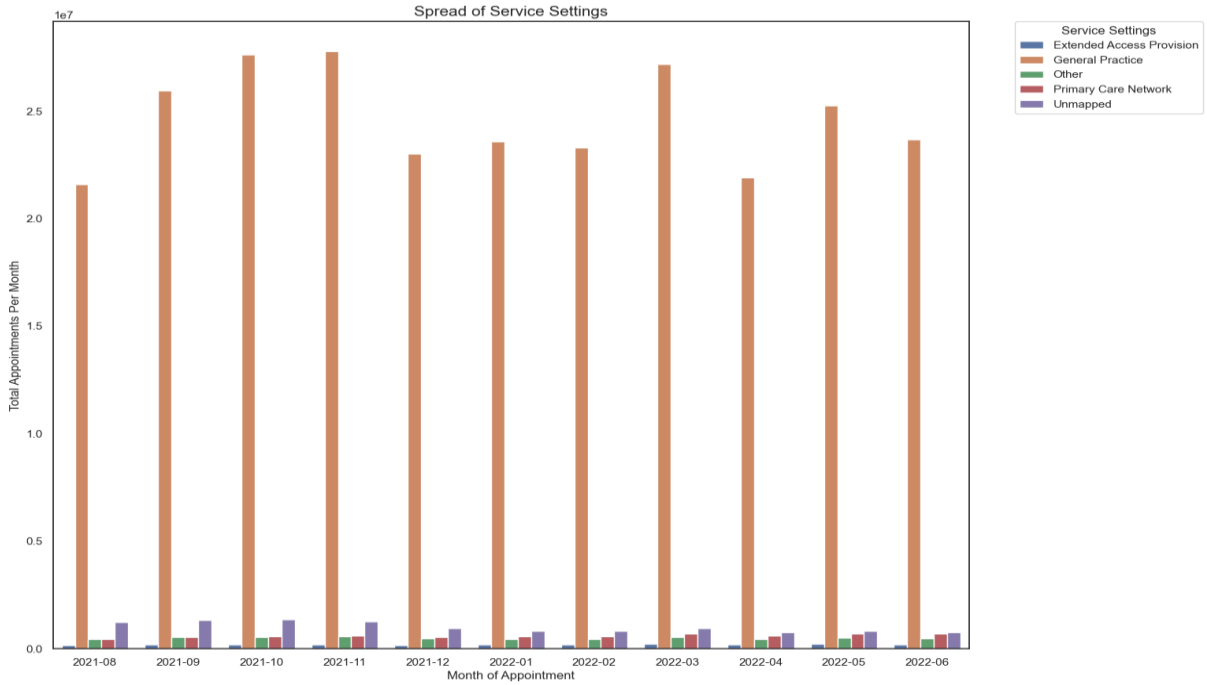

```
# Barplot
sns.barplot(x='appointment_status', y='count_of_appointments', hue='time_between_book_and_appointment',
            data=ar_time, errorbar=None)
plt.title("Time between Booking and Appointment and Attendance", fontsize=15)
plt.xlabel("Attendance of Appointment")
plt.ylabel("Count of Appointments")
plt.legend(title="Time between Booking and Appointment", bbox_to_anchor=(1.05, 1), loc='best')
```



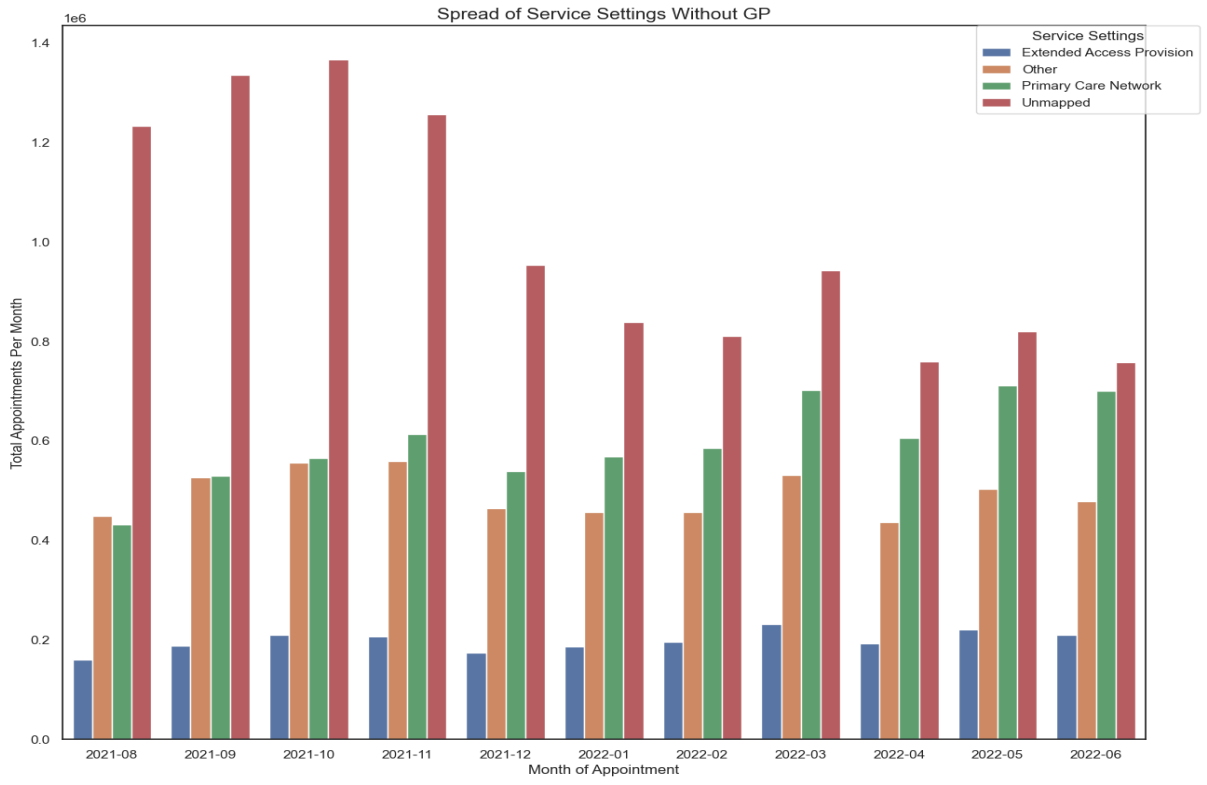

The final consideration is the quality of the data collected. The datasets are not standardised, nor do they start data collection at the same time, providing some incomplete data. Most importantly, they are incomplete in their scope of investigation leaving too many rows with unmapped data or inconsistent mapping, which severely compromises the quality of the data and analysis. As such, the analysis has not progressed to forecasting trends as the quality would have been inconsistent and/or unconvincing.

Comparison of bar charts showing quantity of ‘Unmapped’ category.

```
# Create a boxplot to investigate the spread of service settings.
sns.barplot(x='appointment_month', y='count_of_appointments', hue='service_setting',
            data=nc_sub)
plt.title("Spread of Service Settings", fontsize=15)
plt.xlabel('Month of Appointment')
plt.ylabel('Total Appointments Per Month')
plt.legend(title="Service Settings", bbox_to_anchor=(1.05, 1), loc='best', borderaxespad=0)
plt.show
```




```
# Create a boxplot to investigate the service settings without GP.
sns.barplot(x='appointment_month', y='count_of_appointments', hue='service_setting',
            data=nc_sub[nc_sub.service_setting != 'General Practice'])
plt.title("Spread of Service Settings Without GP", fontsize=15)
plt.xlabel('Month of Appointment')
plt.ylabel('Total Appointments Per Month')
plt.legend(title="Service Settings", bbox_to_anchor=(1.05, 1), loc='best', borderaxespad=0)
plt.show
```



## Conclusion

Overall, the NHS capacity is enough to cover the network but better support to GP practices can be given:

- ‘Telephone’ appointments can support the GP during busy periods.
- ‘Other practice staff’ can support the GP during busy periods.
- Scheduling in the later days of the week can spread the appointments more evenly.
- Reducing the time between booking and appointments can help reduce the number of missed appointments and make the network more efficient.