DA201 Assignment Report
by Jena Shubotheesh

Word count: 1100 without headers and title

Importing and exploring data

While analysing the datasets, I found a total of 106 unique locations. Given that many locations, there are certain to be disparities in appointment data. The value_counts function gave an ordered list of locations by records. We can see 2 parts of London appear in North East and North West. This is to be expected given the denser population within London.

Note that there are a significant number of 'Unknown' under appointment status. This should be noted since it shows a clear lack of information on whether patients did or didn't attend appointments and could reflect a huge amount of unnecessary costs. It also hinders out upcoming analysis.

Initial insights simply showed highs and lows, but this is explored more later.

Analysing the data

Using min and max functions on ad, we can see that appointments were scheduled between 01/12/2021 and 30/06/2022. The national categories dataset goes between 01/08/2021 and 30/06/2022. Both datasets end at the same date, but the national categories dataset goes back an extra 4 months.

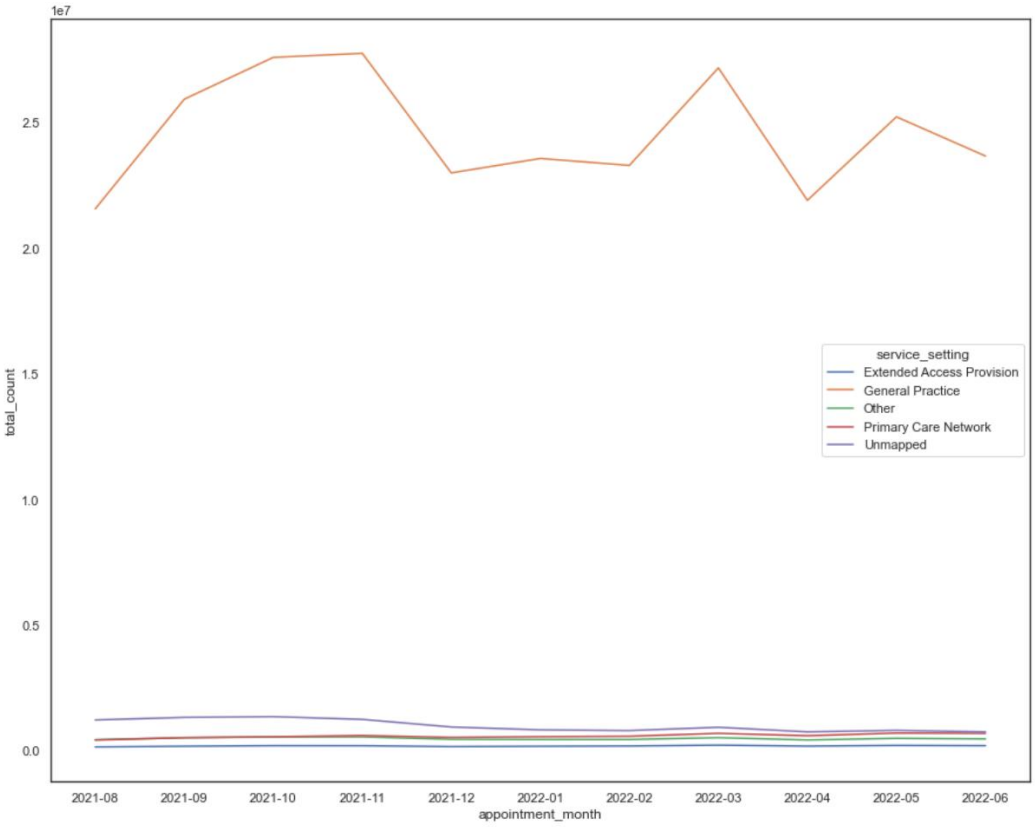
With the groupby function, we can look at the number of appointments per month. We see that 11-2021 had the most appointments with 30,405,070. Just behind it was October of the same year and September was 4th highest. There was consistently high demand for services during that period. Interestingly the lowest month of appointments was August 2021 suggesting a sudden spike after.

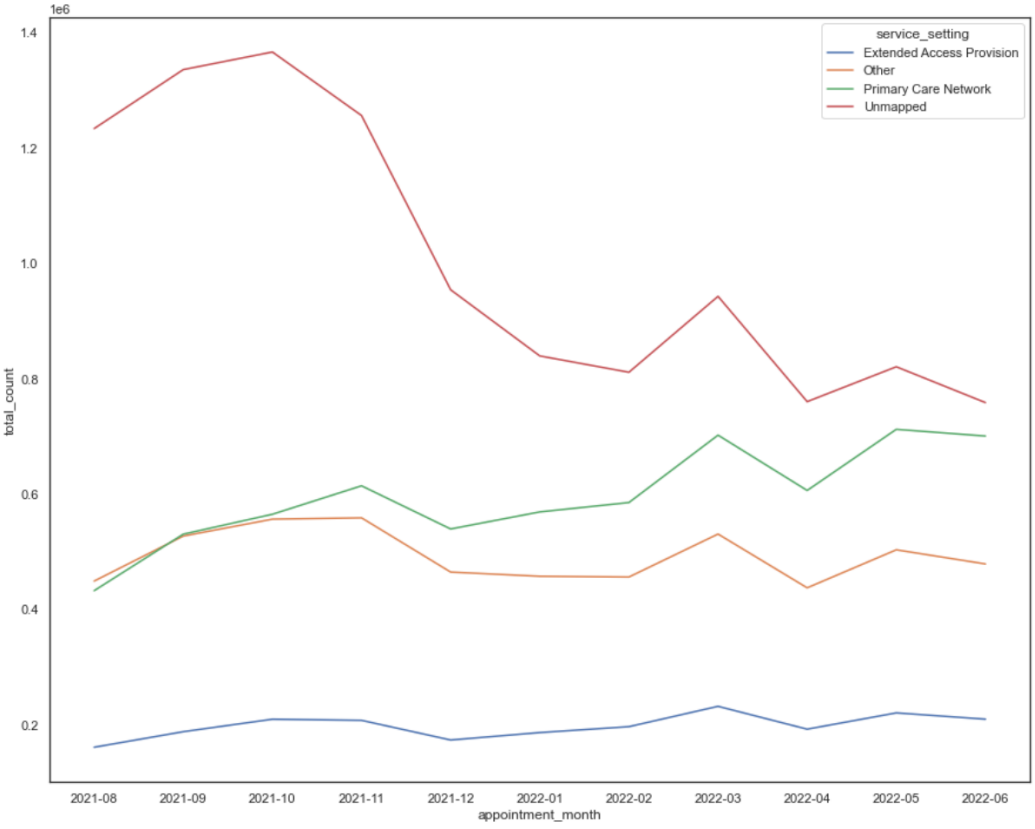
The count function lets us look at total number of records per month. We see March 2022 have the highest with 82,822. While that corroborates with the high count of appointments in that month, generally there isn't any correlation between the two subsets.

Visualising and identifying initial trends

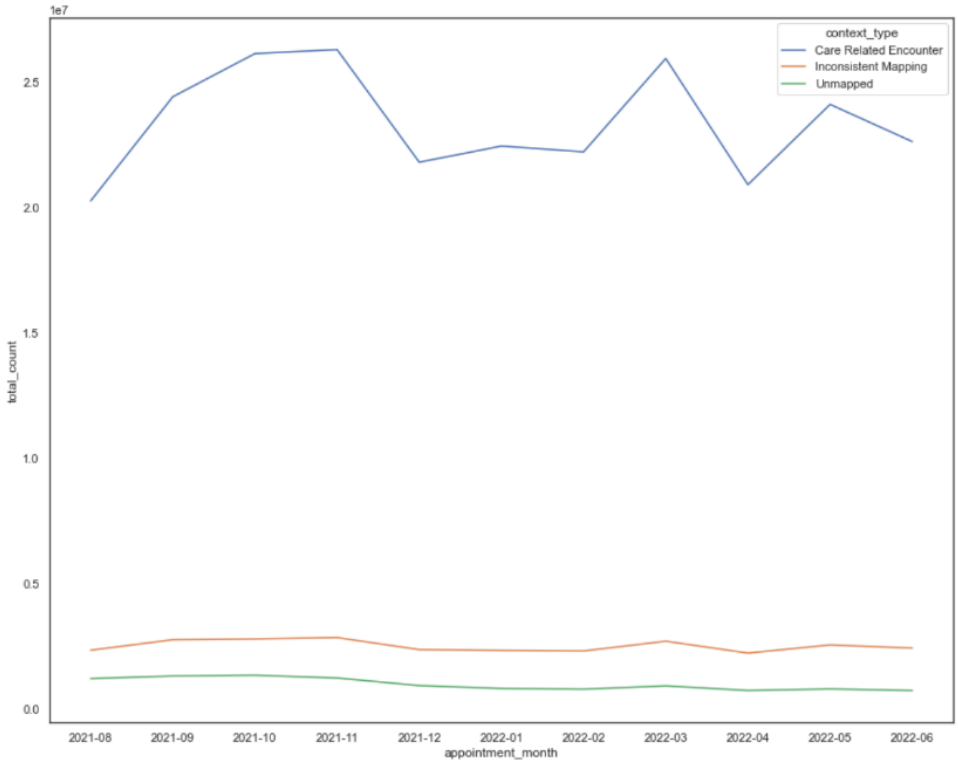
When plotting service settings over months, there is an obvious dominance of 'General Practice' compared to others. We generally see a fluctuation over this period of time with 08-2021 and 06-2022 being at similar points of just over 20,000,000 appointments.

To look at the other service settings further, I replotted the data excluding 'General Practice'. Here we see 'Unmapped' being highest by a large margin in 08-2021. However, we see a general decline here over time, suggesting better recording of service settings. 'Primary Care Network' had a slight increase of around 200,000 while the rest remain the same.

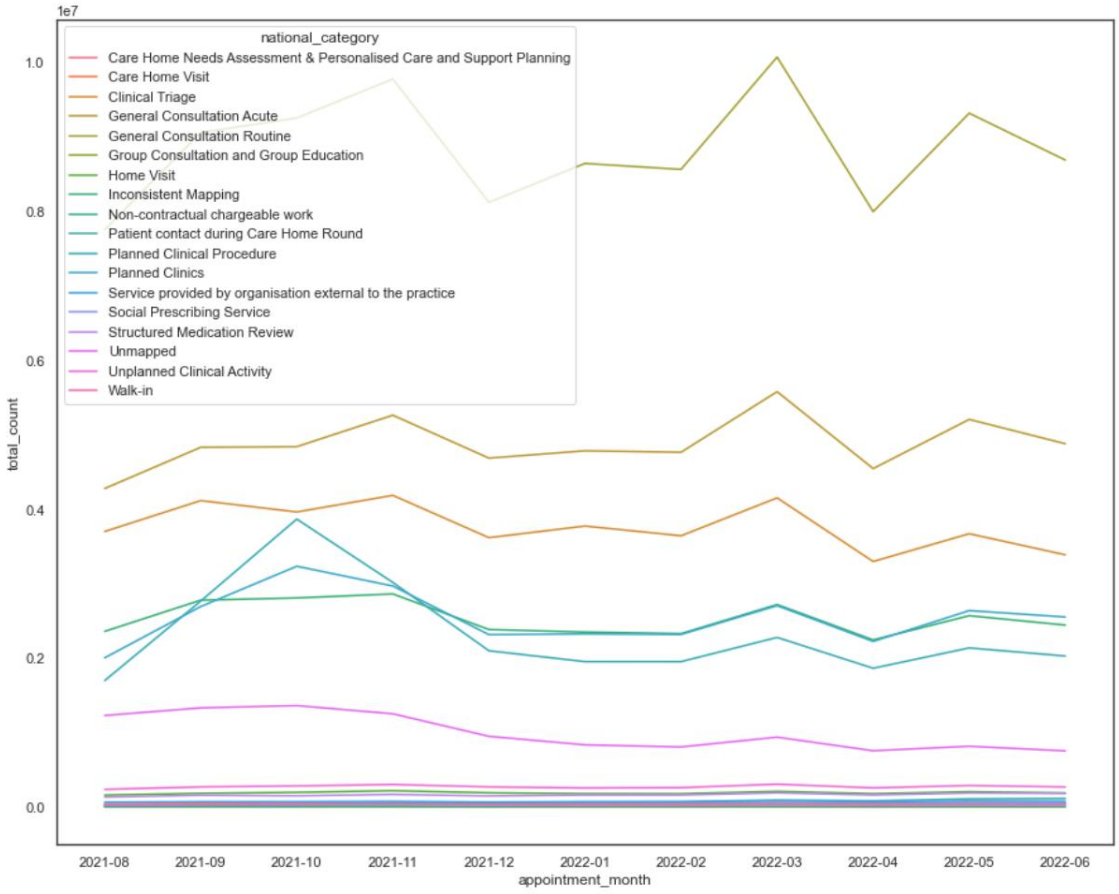

1e7
2.5
2.0
1.5
1.0
0.5
0.0
total_count
service_setting
Extended Access Provision
General Practice
Other
Primary Care Network
Unmapped
2021-08
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
appointment_month

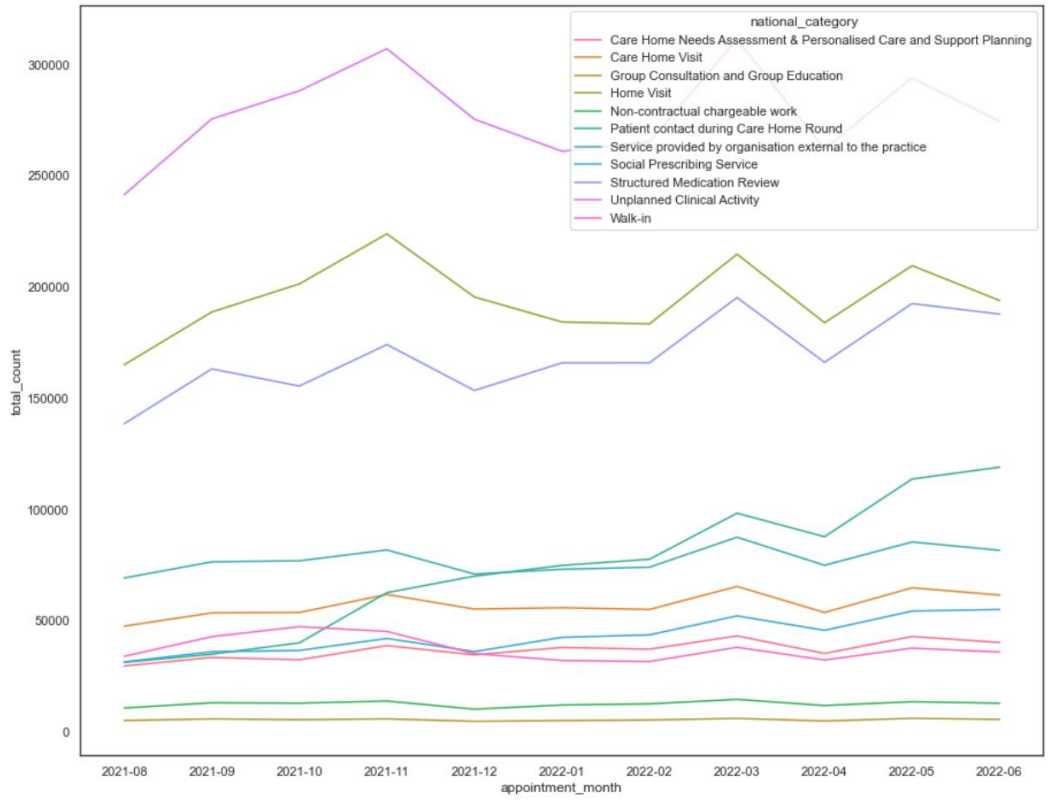

1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
total_count
service_setting
Extended Access Provision
Other
Primary Care Network
Unmapped
2021-08
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
appointment_month

A similar situation arose for context types. 'Care Related Encounter' is highest with around 20,000,000. The low level of inconsistent or unmapped values suggests this graph is quite effective.

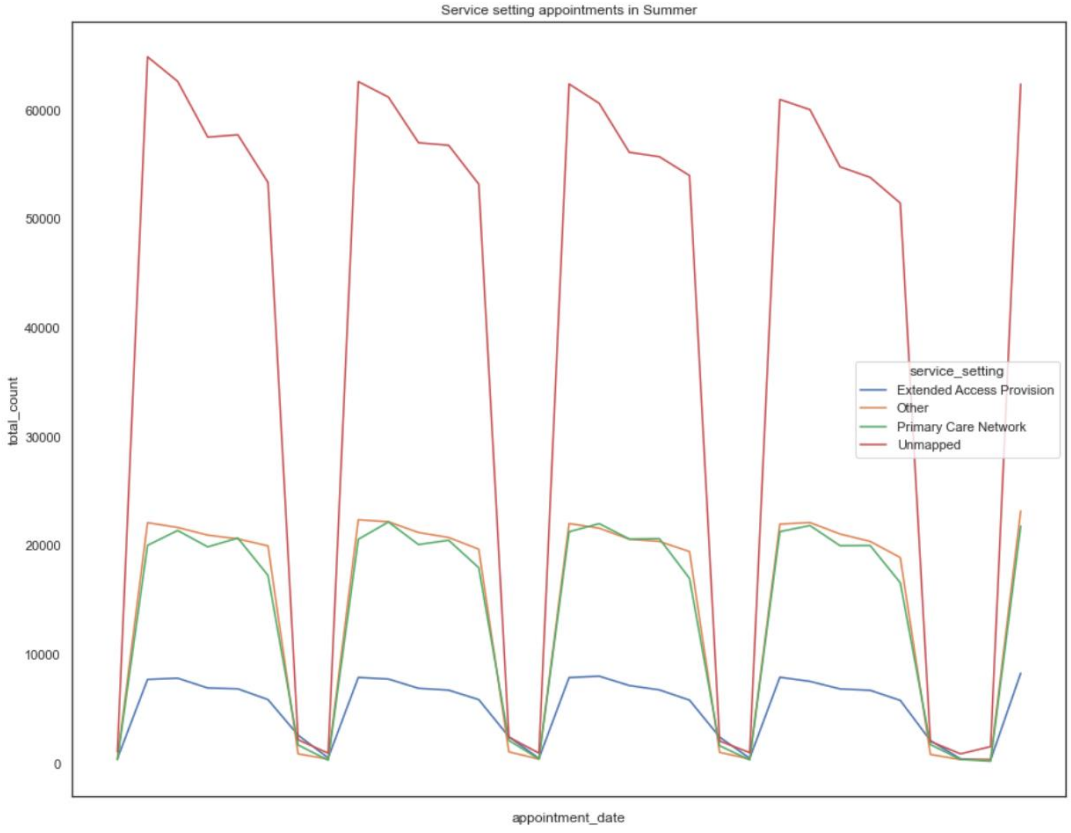


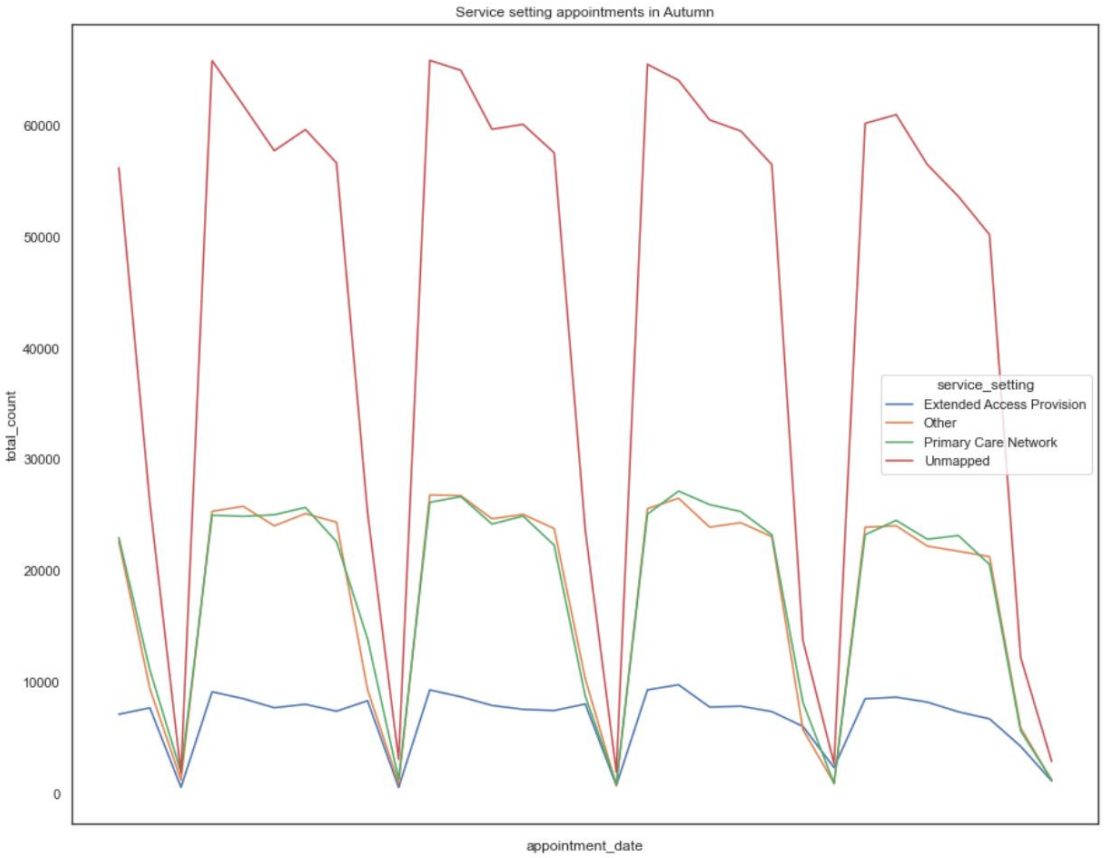
National category has a lot more lines, with 'General Consultation Routine' having the highest. Planned clinical procedure had a large spike in 10-2021. Removing some of the higher value lines showed the other categories clearer. While all other lines moved in parallel, there was a great increase in 'Patient contact during Care Home Round' over time. Perhaps this is due to COVID and its greater impact on older victims. Comparing to the former diagrams, we can see that the pattern is similar, with a consistent spike in 03-2022.

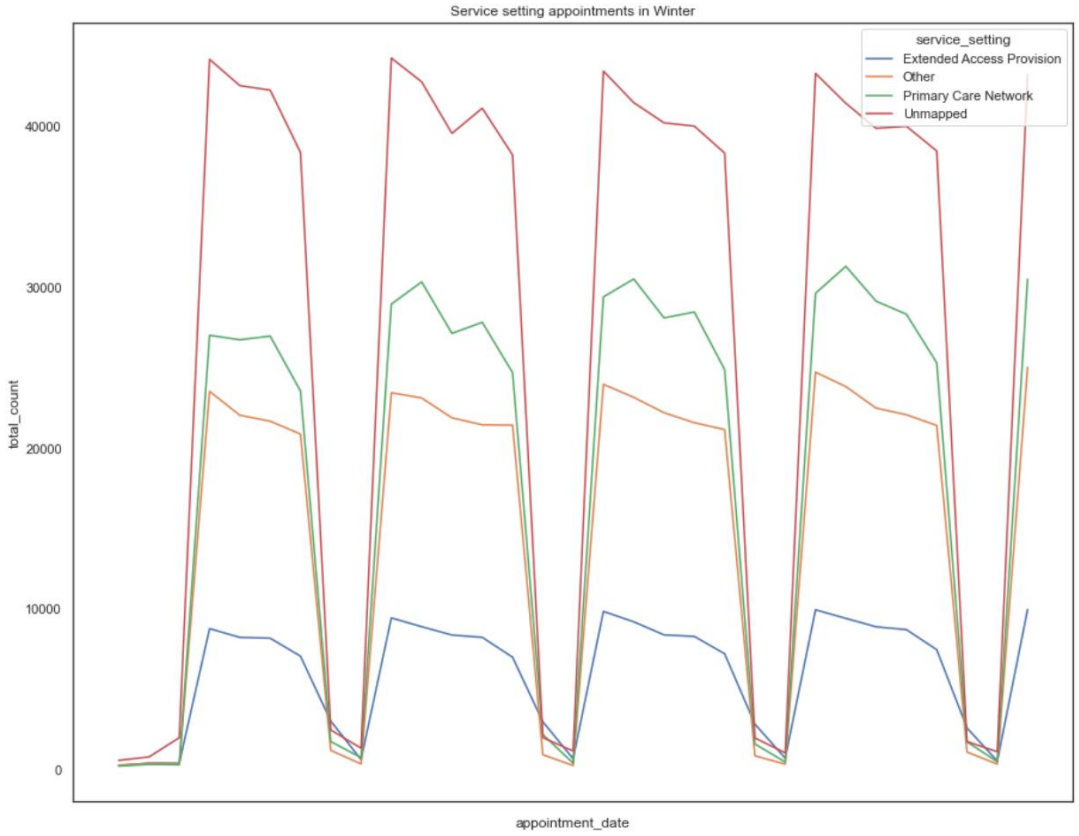


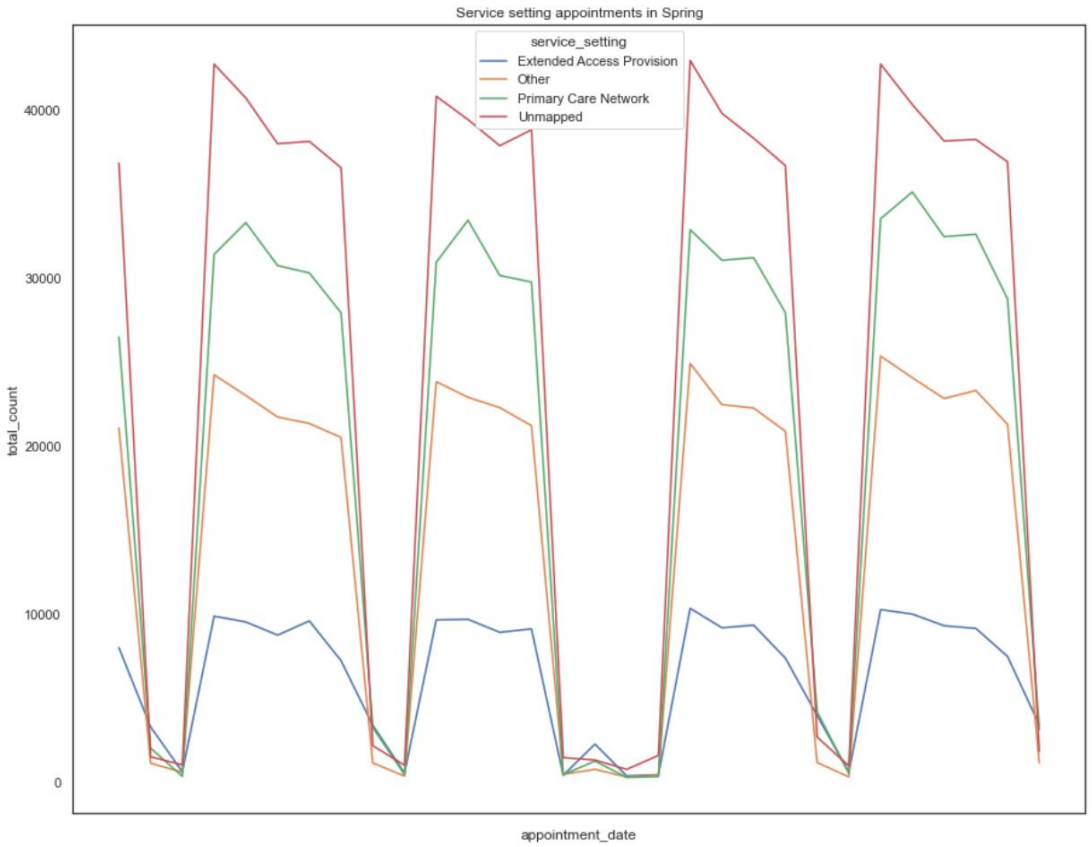


Looking seasonally, we see most appointments are 'Unmapped' for all seasons. 'Primary Care Network' started at the same place as 'Other' in Summer but continued to increase in gap over each season. Summer and Autumn have higher peaks of just over 60,000 while the following two seasons are around 40,000. The decline in unmapped is a good symbol of improved recording.



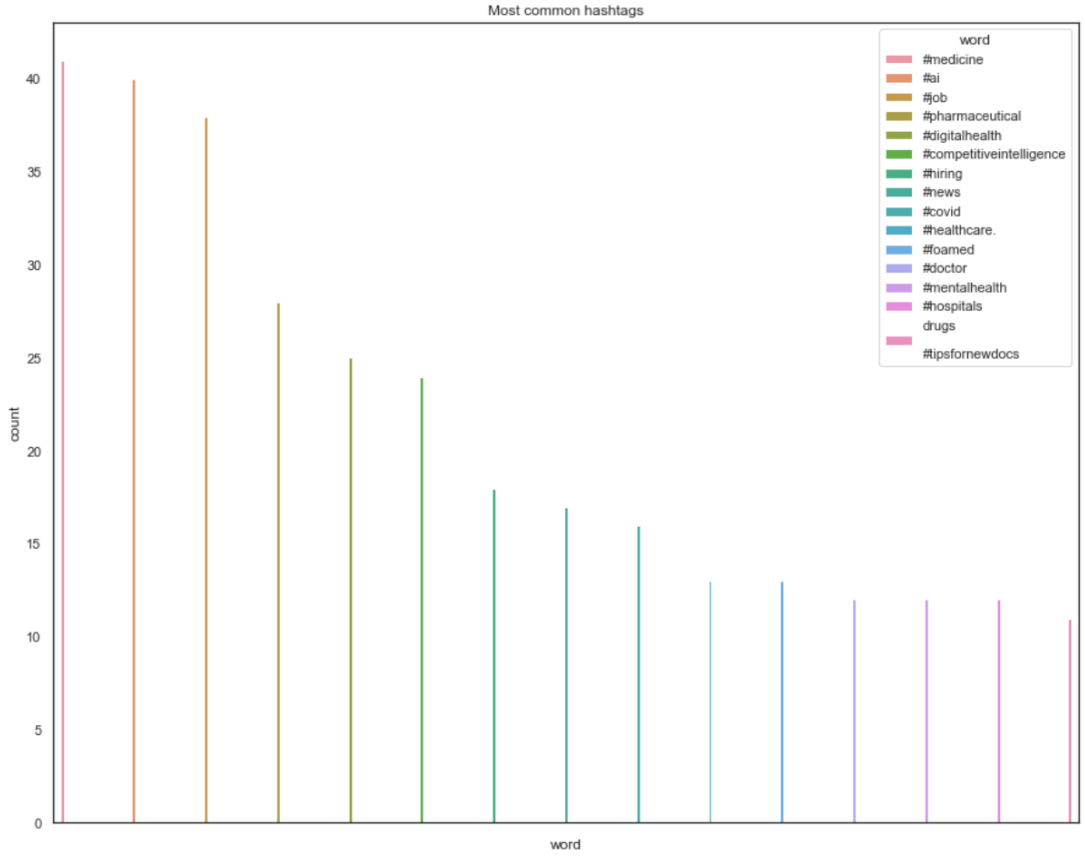

Service setting appointments in Autumn
total_count
60000
50000
40000
30000
20000
10000
0
appointment_date
service_setting
Extended Access Provision
Other
Primary Care Network
Unmapped


Service setting appointments in Winter
total_count
40000
30000
20000
10000
0
appointment_date
service_setting
Extended Access Provision
Other
Primary Care Network
Unmapped



Analysing the Twitter data

When analysing the twitter data, there were some immediate insights that would be found. The hashtag 'healthcare' was most dominant with 716 counts followed by 'health' with 80. This made it imperative to simplify this dataset to be able to produce a diagram that shows a small but versatile range of hashtags rather than many similar ones. Ideally, we would be able to look at all hashtags but given limited space and time we had to simplify.



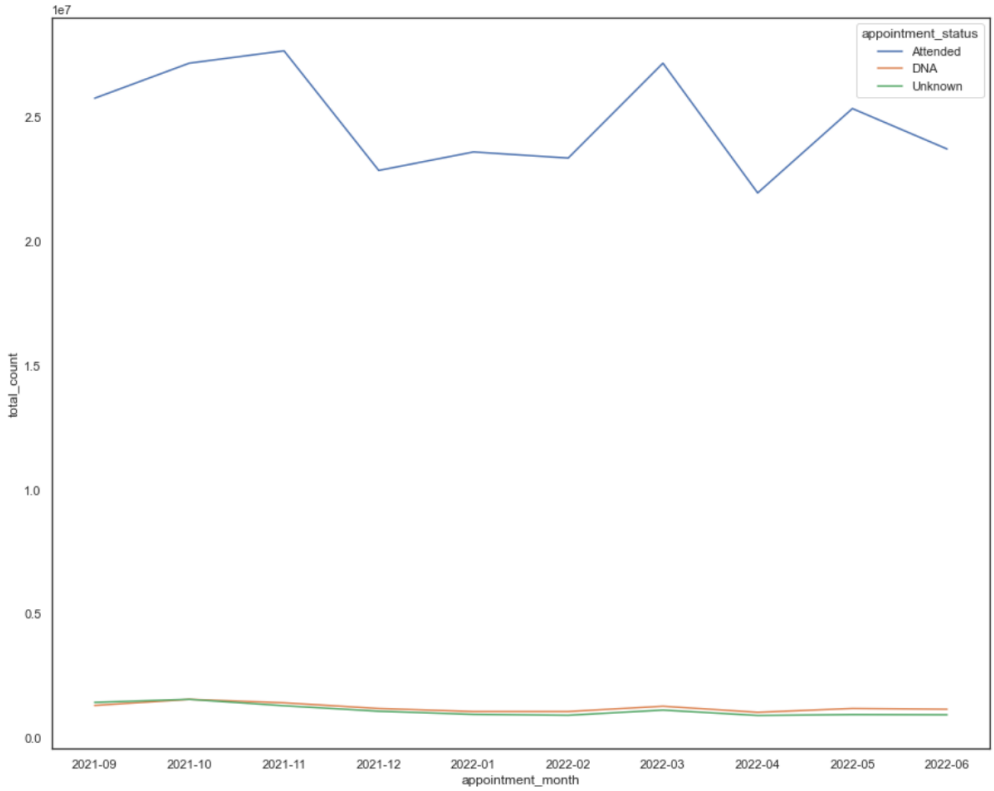
I therefore dropped a large majority of words before creating a bar plot. Note that the words kept are what I felt represented a wide variety of topics. A different set of words may have yielded different insights. From my dataset, we see #medicine having the highest number of uses with 40. Perhaps this symbolises popularity across medical products. There is also a high number of tweets including technology related words such as #ai and #digitalhealth. This suggests a potential interest in that side of the medical field. #covid is only around the middle of our diagram. Suggesting it may not be the dominant cause of appointments

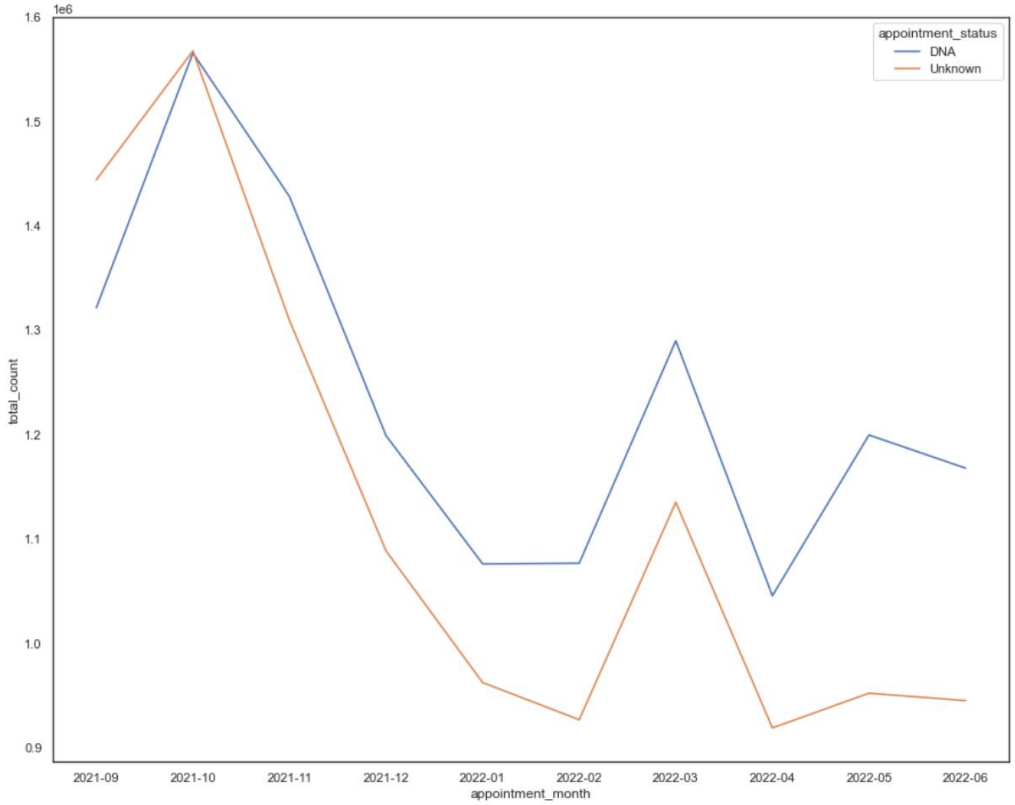
Making recommendations

Finally, when looking at monthly appointments and utilisation, we see a great fluctuation in numbers but identical lines. The highest point is 11-2021 with 04-2022 being the lowest. There is generally a wide amount of fluctuation in between these values. In the utilisation diagram, I have added an extra line representing the maximum accommodation of 1,200,000. As we can see, there are no months where the NHS have exceeded this, showing an adequate level of staff for demand.

It may even be beneficial to reduce the capacity to around 1.05 million daily since our line doesn't exceed that. Note that we divided by 30 to get daily values and this may not be accurate - there may be significant spikes in daily utilisation that is to be expected with the sporadic nature of illnesses.

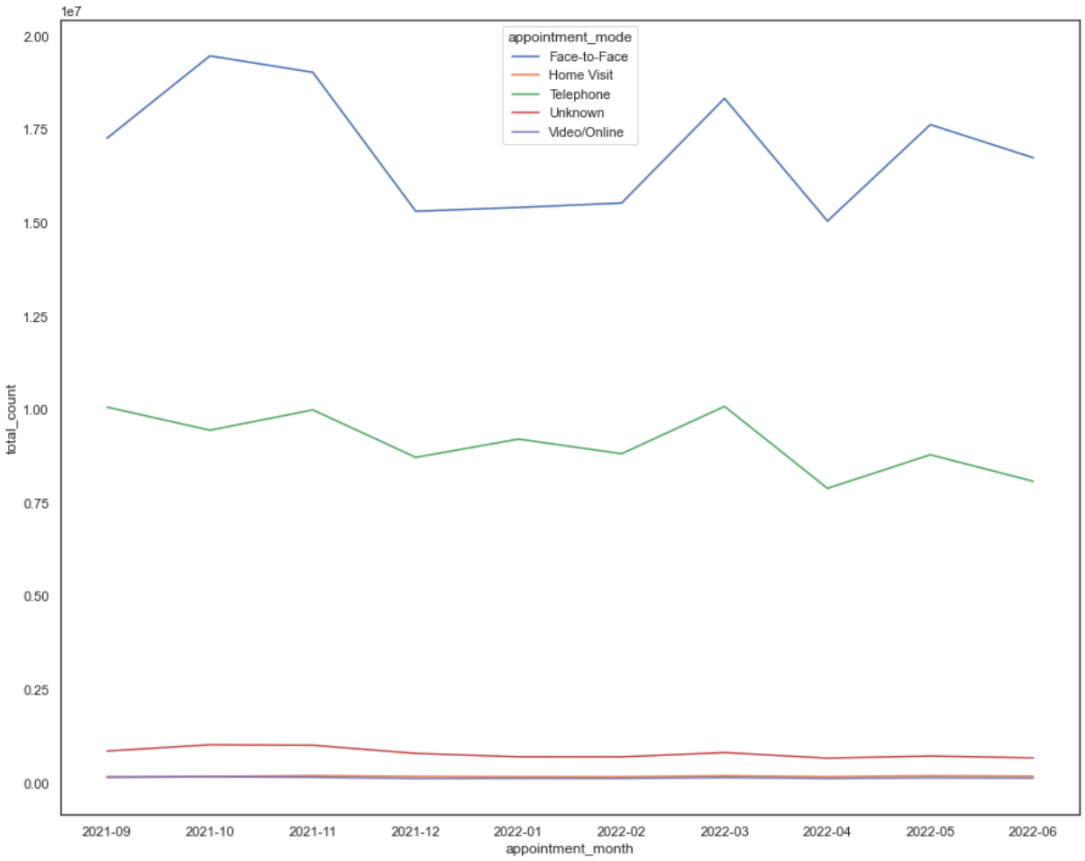
Looking at appointment status, there have mostly been attendance to appointments. Removing this, we see 'Unknown' and 'DNA' following similar trends. There has been an overall decrease, with a slight increase in 03-2022. The gap between the two has increased over this time period.

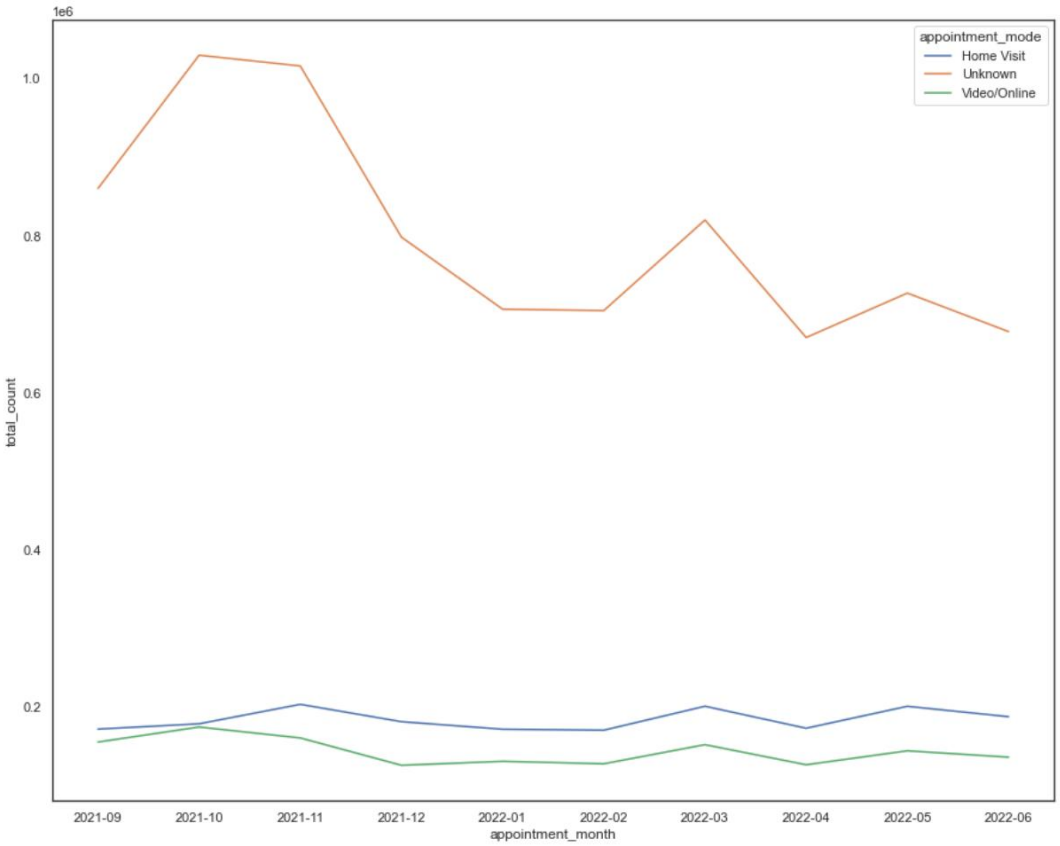

1e7
appointment_status
Attended
DNA
Unknown
2.5
2.0
1.5
1.0
0.5
0.0
total_count
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
appointment_month


1e6
appointment_status
DNA
Unknown
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
total_count
2021-09
2021-10
2021-11
2021-12
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
appointment_month

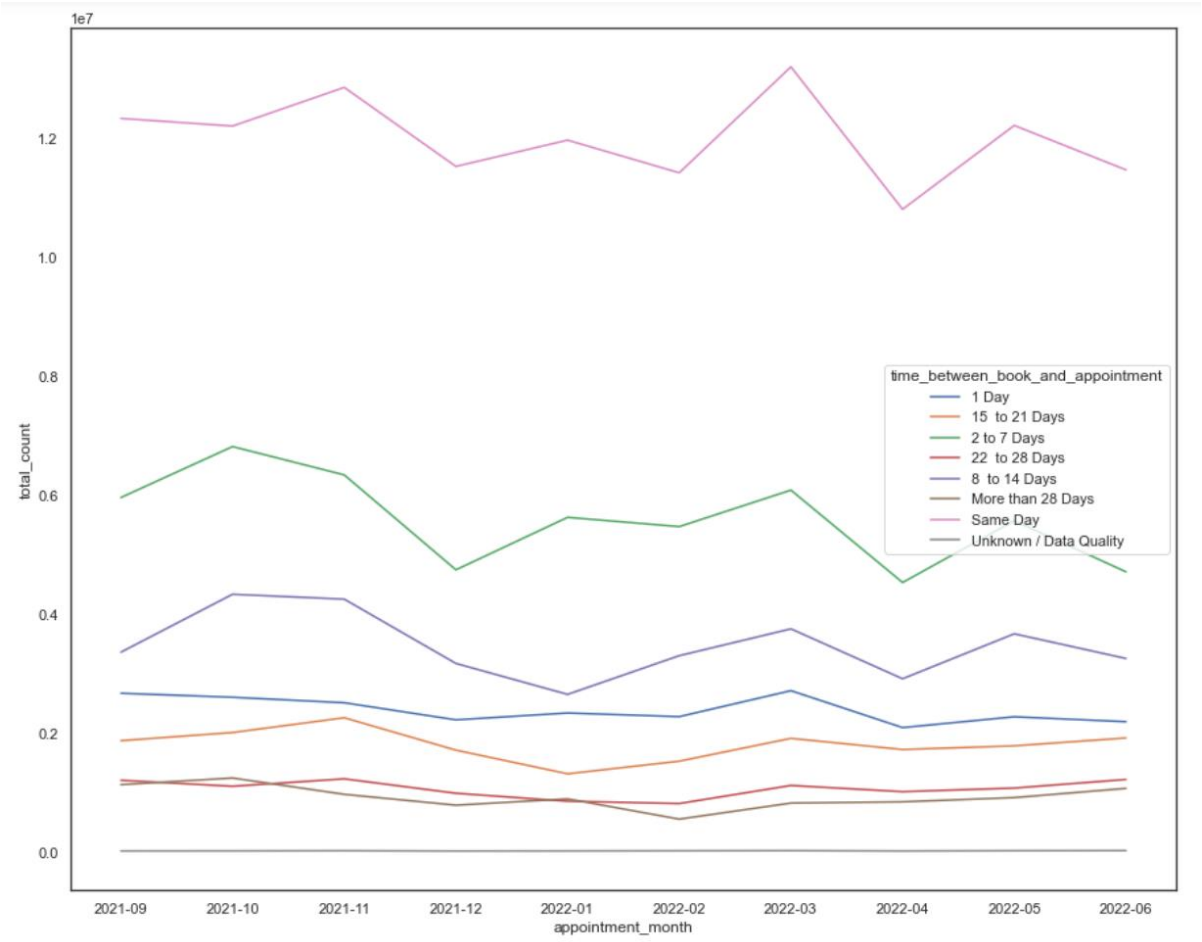
We can see hcp type having low unknown and almost equal GP and Other staff. Other staff represents many roles which could not be specified by NHS when providing this data, so we can not look any further.

Face-to-face and telephone appointments were most popular form of contact, with a large number of 'Unknown' behind. Home visits and video appointments were not popular at all during this period. All fluctuate but start and end at similar points.

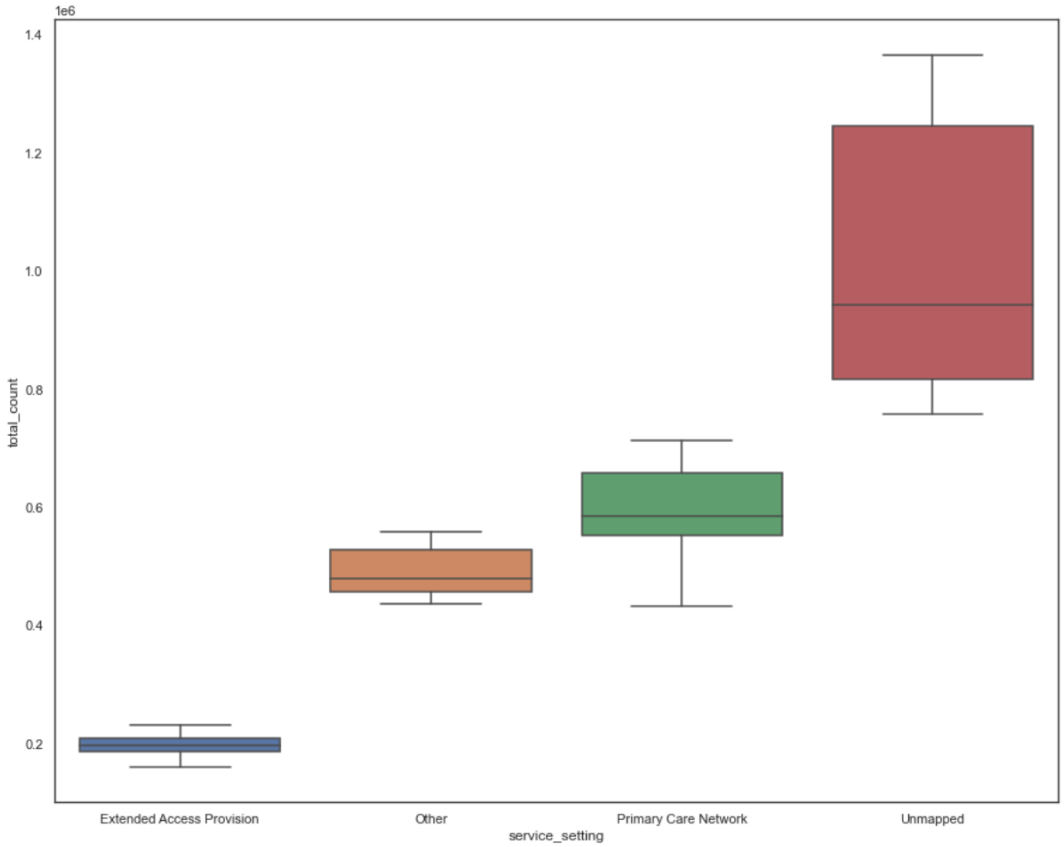




Most appointments were attended on the same day. The lines below generally seem to follow an order of increasing time. Clearly patients who did attend their appointments were able to as soon as possible, suggesting that there was adequate resource to respond to their needs.



Looking at boxplots, we see a wide range on 'Unmapped' counts and the lowest Q1-Q3 range in 'Extended Access Provision'. There seems to generally be a poor amount of service setting data provided in the dataset, suggesting either poor record collecting or another issue.



The metadata suggested that there has generally been poor recording of data and understanding of attendance and this has definitely skewed our analysis and visualisations.