Final Assignment Report

The Business problem: The NHS would like to identify the reason behind the missed appointments throughout the country and therefore reducing enormous financial implications it generates.

The immediate two main questions posed by the NHS are:

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

In order to reflect to answer the broader main questions, further analysis are required and the following investigations have been undergone:

- What is the number of locations, service settings, context types, national categories and appointment statuses?
- The date ranges of the provided datasets and which service settings reported most appointments for the period.
- The number of appointments and records per month
- The monthly and seasonal trends, based on number of appointments for service settings, context types and national categories
- Were there adequate staff and capacity in the networks?
- What possible recommendations does the data provide the NHS?

The data analysis process. The process includes data Ingestion, data wrangling, creating visualisations to identify insights and patterns, determining any outliers in the data are undertaken.

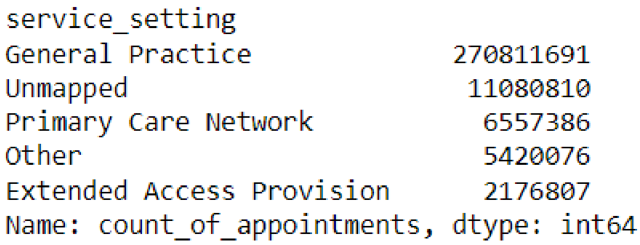
Data ingestion begins with importing the necessary python libraries to run the code, complete the analysis and draw up visualisations. The most important Python libraries that are imported include Pandas, NumPy, Seaborn, Matplotlib, datetime and yaml for twitter analysis. Then the requisite datasets are loaded on to Pandas. 'pd' is used as the alias for Pandas. The Three datasets used for this project are: actual_duration.csv, appointments_regional.csv and national_categories.xlsx. These files are loaded using the respective Pandas functions like read_csv() and read_excel(). Various data sense checking and statistical analysis are carried out on the imported datasets. I would like to note that, in spite of the common key 'icb_ons_code' on all datasets, any merge were not done as the dataset was not in a normalised status (duplicates for primary key constraints). The most common sense checking process, being, displaying the shape, dtypes, head(),tail(), describe() and info() of the dataset are carried out. The Three datasets are also checked for any missing data and confirmed to be none.

In order to find out the number of locations, I use 'sub_icb_location_name' column in 'nc' dataframe and value_counts(). Ther are 106 locations in total. Then the top five locations are identified , which are :

```
Top five locations: NHS North West London ICB - W2U3Z              1300
7
NHS Kent and Medway ICB - 91Q                  12637
NHS Devon ICB - 15N                            12526
NHS Hampshire and Isle Of Wight ICB - D9Y0V    12171
NHS North East London ICB - A3A8R              11837
```

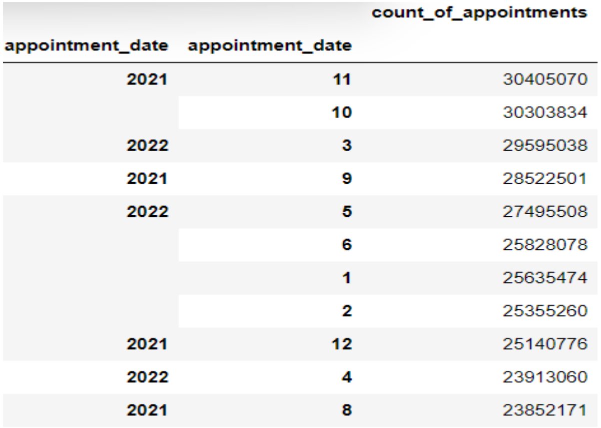
Then the total number of service settings(5), Context types(3), national categories(18), appointment statuses(3) and appointment modes(3) are determined. The 3 main dataframes used are ‘ad’ for loading ‘actual_durations.csv’, ‘ar’ for ‘appointments_regional.csv’ and ‘nc’ for ‘national_categories.xlsx’. Minimum and maximum dates are calculated for each dataset.

The service setting with the most appointments in the NHS North West London ICB - W2U3Z, with in the specified date ranges- between 01/01/2022 to 01/06/2022 are calculated.



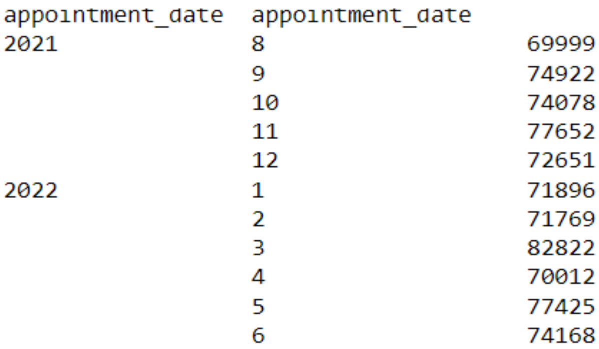
```
service_setting
General Practice            270811691
Unmapped                     11080810
Primary Care Network          6557386
Other                         5420076
Extended Access Provision     2176807
Name: count_of_appointments, dtype: int64
```

Then the month with highest appointment number were calculated:



		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

The total number of records per month was also calculated as:



```
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
```

In most of the cases the split-apply-combine process is maintained and groupby() with aggregation functions(e.g. agg()) are used appropriately with ‘sum’, ‘count’, ‘mean’ etc. and sort_values() functions to complete the aggregations.

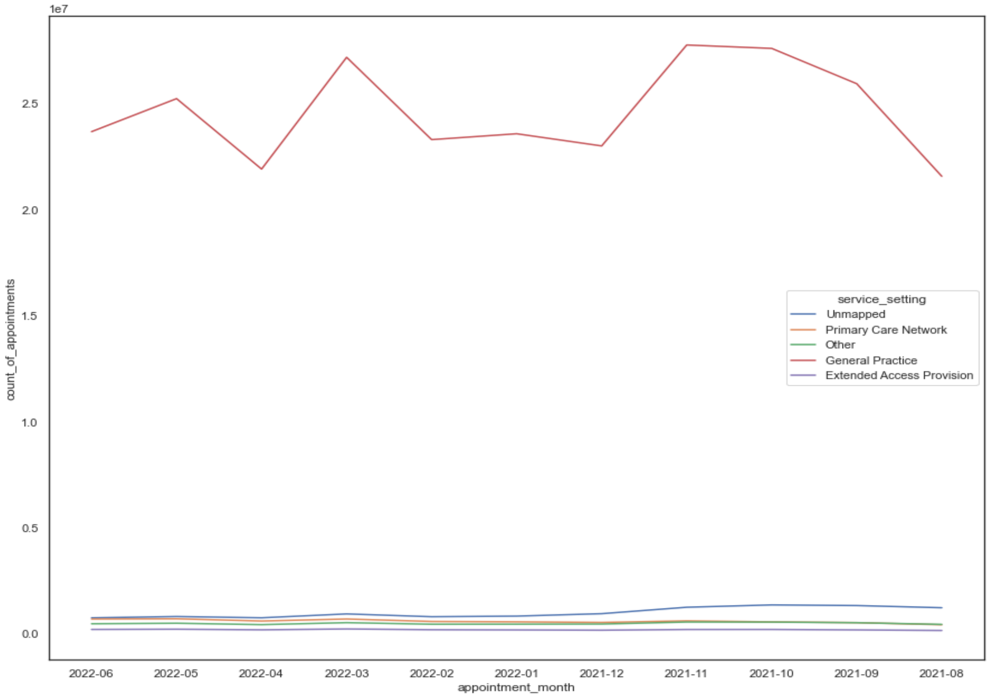
Visualisations overview

In order to answer the business questions, various charts have been plotted.

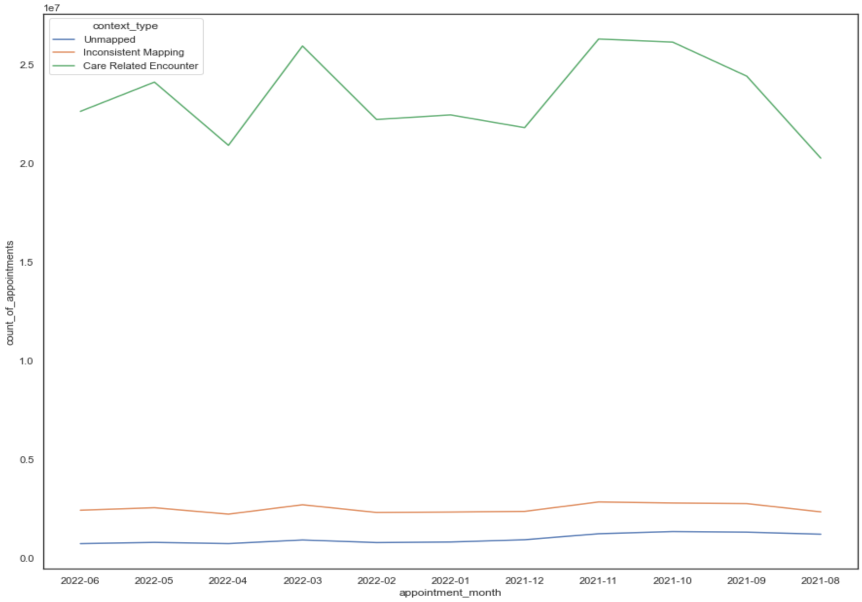
Appointments per month for service setting.

As evident, General Practice had the most appointments, followed by Unmapped setting. The peak of the appointment are around the November 2021 and has been slightly decreasing as of

06/2022. The graph also shows the number of appointments difference within the service settings. The Unmapped appointments is the key area which needs further investigations.

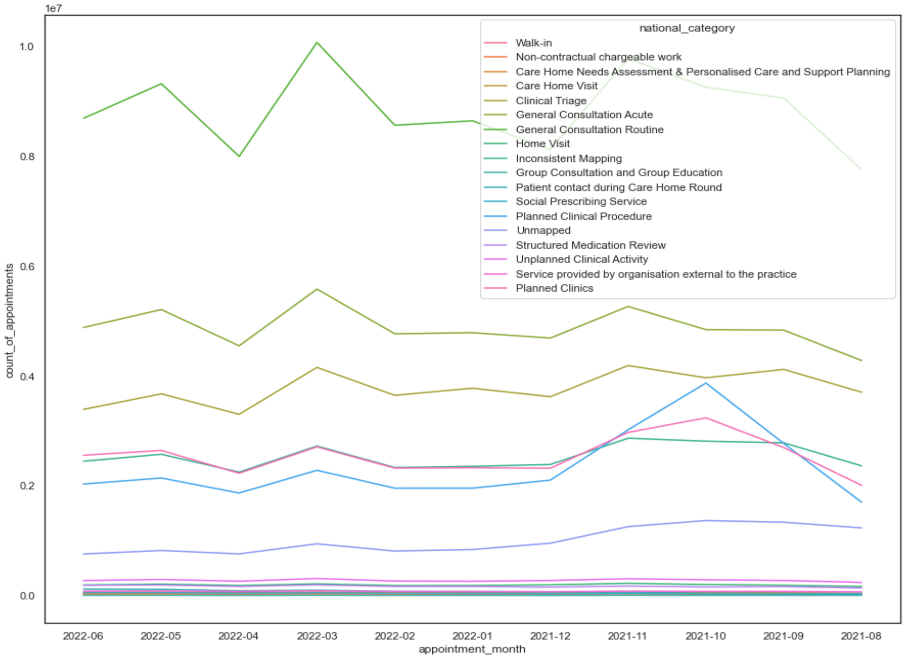


Appointment for context types.



As observed, Care Related Encounter has the maximum appointments, followed by inconsistent mapping and unmapped types. The peak around November 2021. There are significant amount in Inconsistent mapping and unmapped types , which again needs to be looked into.

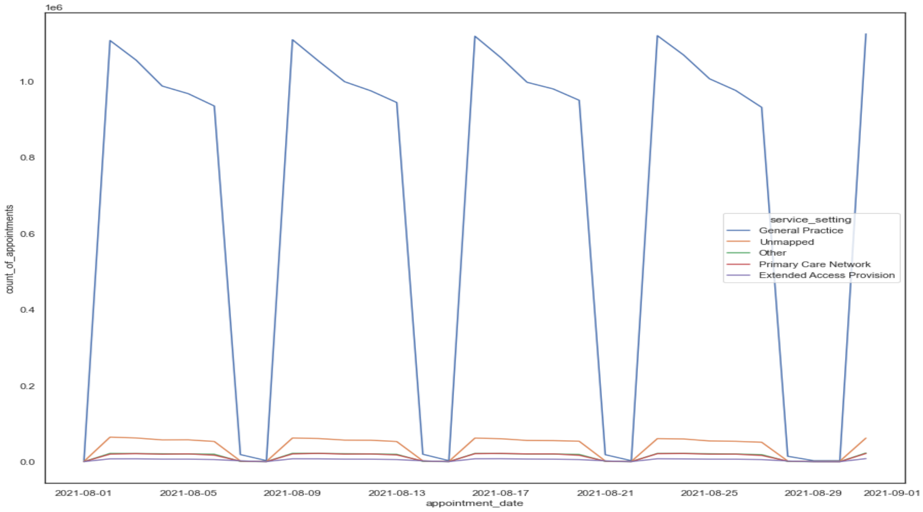
Appointments for national categories:



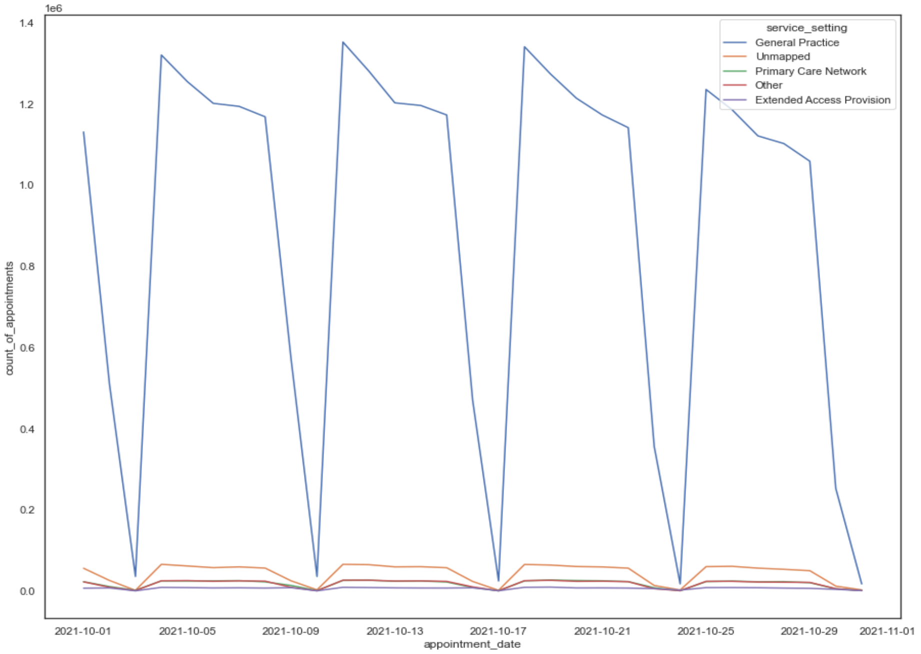
General consultation routine have the most appointments followed by General consult acute and clinical triage. Inconsistent mapping, unmapped have significant records which , as in the previous chart categories. The peak appointments were in March 2022 , slightly declining as of June 2022.

Seasonal appointments including Summer21, Autumn 21, Winter 22, Spring 22.

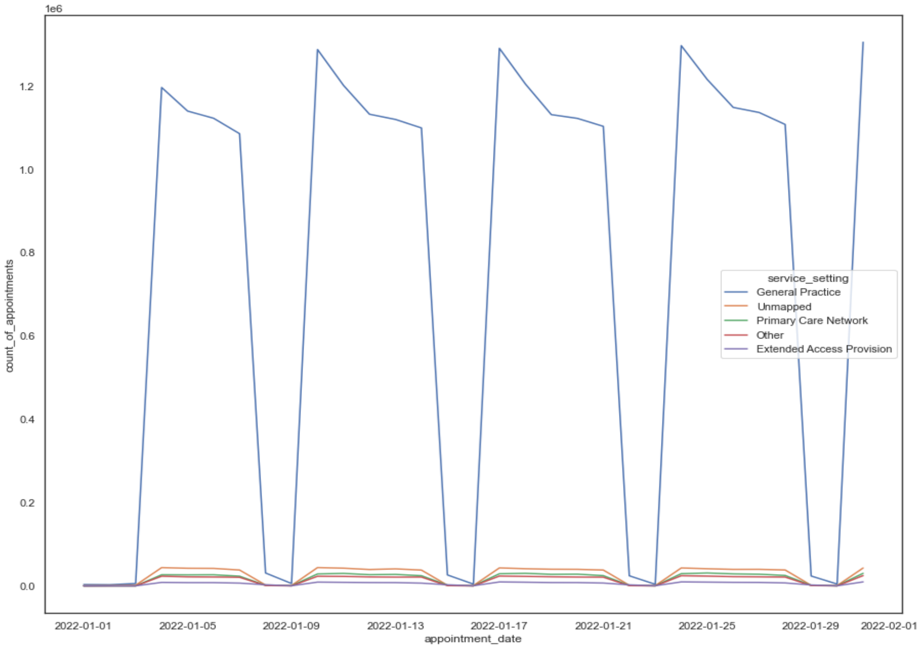
Summer 21



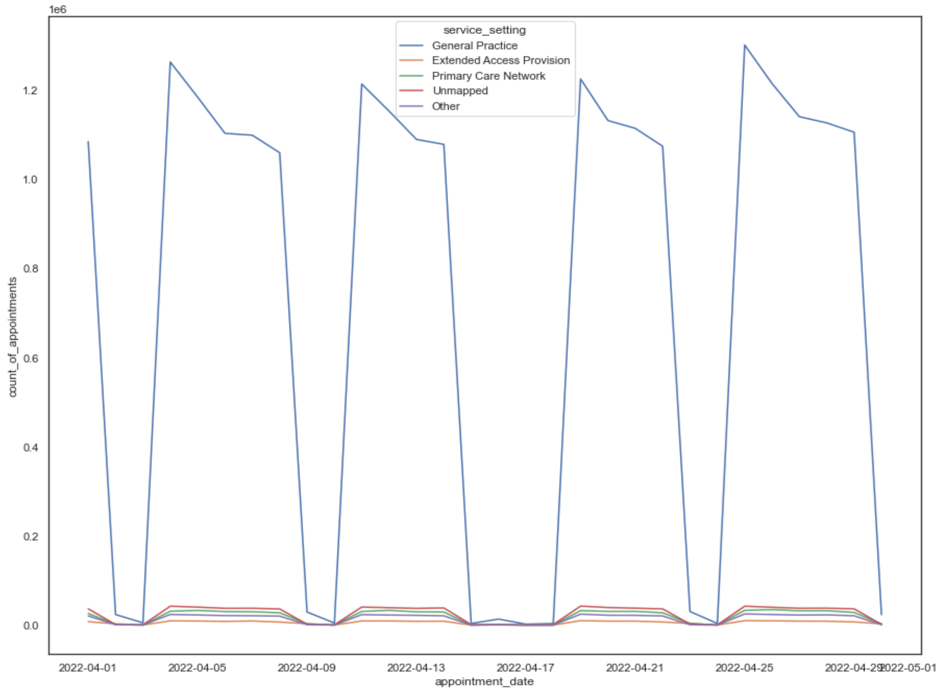
Autumn 21



Winter 22



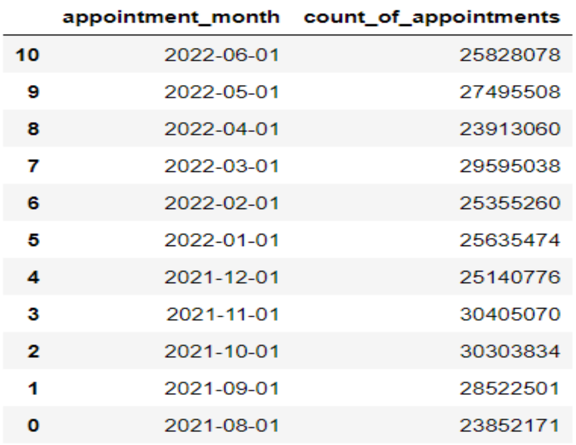
Spring 22



As it stands, General practice have the most appointments with a similar pattern across the seasons but significantly followed by Unmapped type.

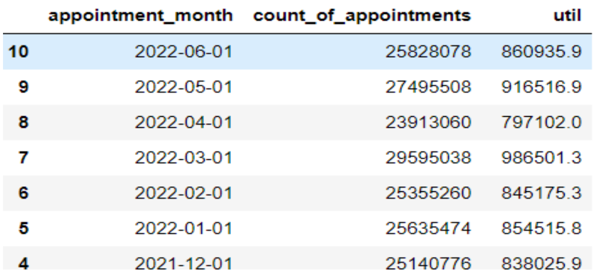
Should the NHS start looking at increasing staff levels?

Various aggregated have been produced to support. Appointments per month is calculated as per below data table.



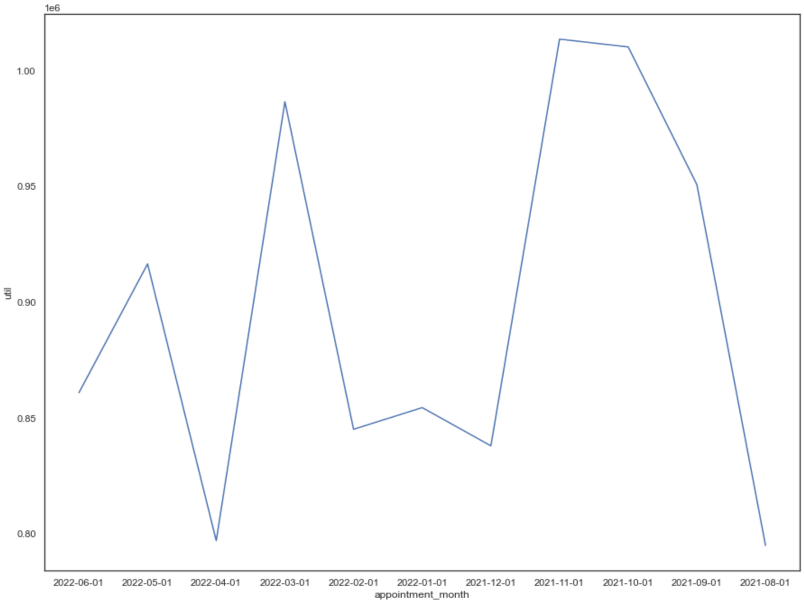
	appointment_month	count_of_appointments
10	2022-06-01	25828078
9	2022-05-01	27495508
8	2022-04-01	23913060
7	2022-03-01	29595038
6	2022-02-01	25355260
5	2022-01-01	25635474
4	2021-12-01	25140776
3	2021-11-01	30405070
2	2021-10-01	30303834
1	2021-09-01	28522501
0	2021-08-01	23852171

To understand utilisations, a new calculated column ‘util’ is added , rounded to 1 decimal place.



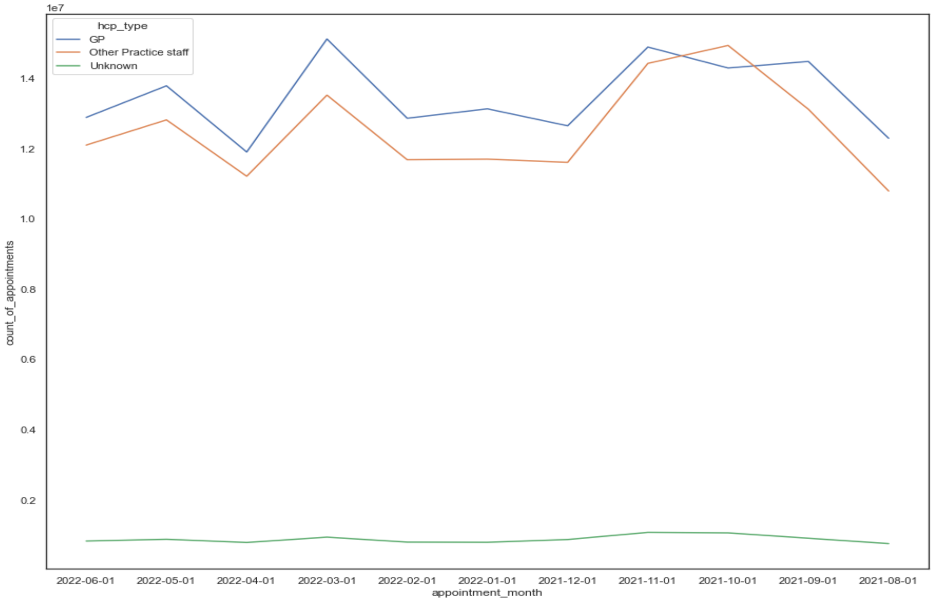
	appointment_month	count_of_appointments	util
10	2022-06-01	25828078	860935.9
9	2022-05-01	27495508	916516.9
8	2022-04-01	23913060	797102.0
7	2022-03-01	29595038	986501.3
6	2022-02-01	25355260	845175.3
5	2022-01-01	25635474	854515.8
4	2021-12-01	25140776	838025.9

To calculate and plot the utilisatiions,



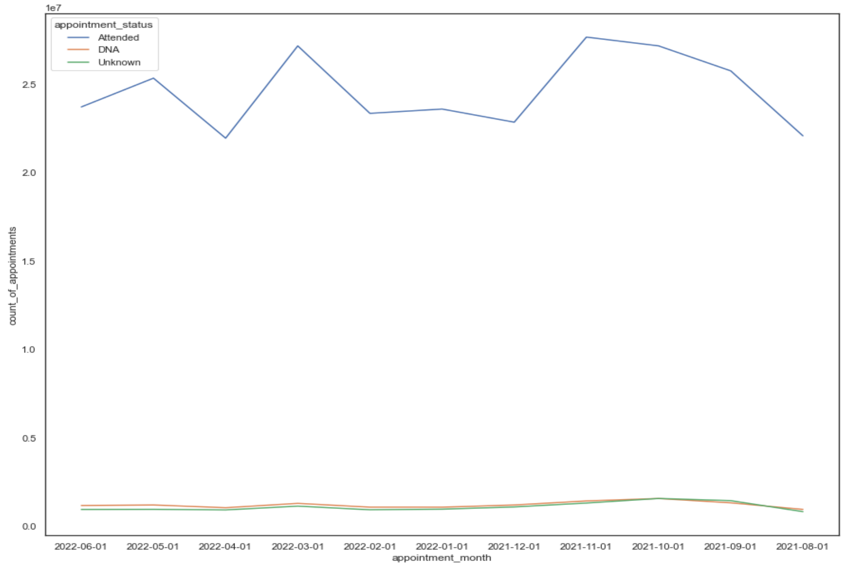
Utilisations seems to be peaking around November 21, again back in Jan 22, but as it stands is slightly declining as of June 22.

To determine how healthcare professional types differ over time by Line plot



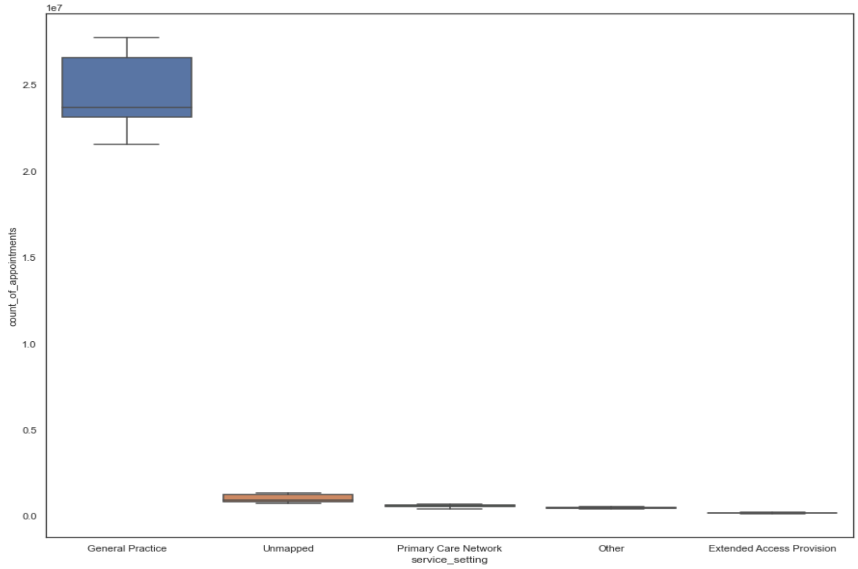
As seen above, the health care professional which has most appointments are GP and other practice staff and also ‘Unkown’ type is present which fails to represent which professional has attended. Both the types have a similar pattern with the maximum appointments around March 2022.

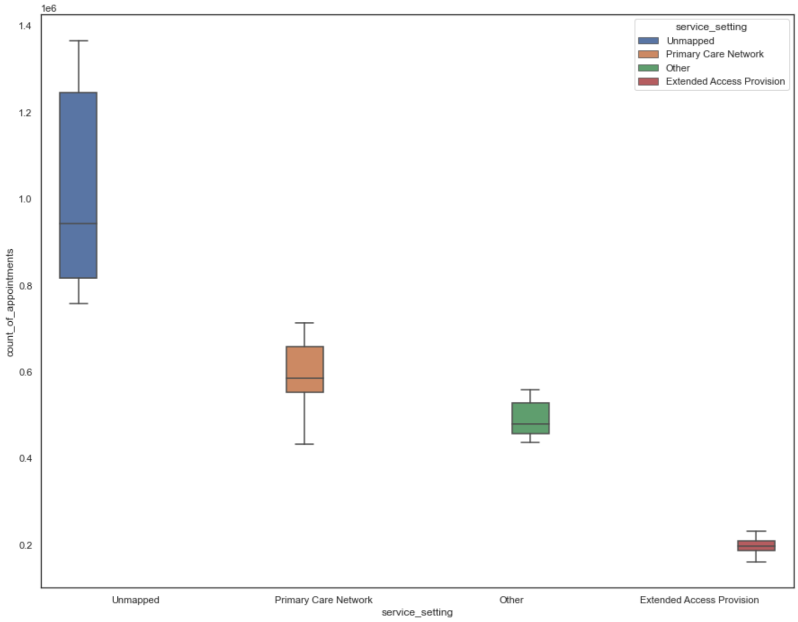
Are there significant changes in whether or not visits are attended?



I think yes, as we can see the most appointments constitute the attended appointments but will have impact if DNA rates goes up.

Finally , comparisons of service settings are done. The second chart is drawn by excluding the GP.





I am using Boxplots in these two charts as they displays the distribution of numerical data and skewness based on 5 points: min,Q1,median, Q3, max.

Boxplot gives a fair indication of how the data is spread out. As observed GP has most appointments followed by the Unmapped and then primary networks. In the second chart GP is removed for comparison, again Unmapped becomes the most recorded category , followed by PCN.

Patterns observed.

I think the networks utilisations are within the framework but the observed pattern consistently of Inconsistent mapping and unmapped among Context types, Unmapped in the service setting categories, Unknown category in Healthcare professional draws a pattern of error in recording the data capture. This could be because of data recording procedure incompatibility or data entered or identified wrongly in the systems. These categories need to be investigated further to improve the data representation and quality. All the questions were worked out using the data analysis phase.

Recommendations:

Ideally improve the data capturing methodologies across all NHS systems. One way this could be achieved is by using a universal system across all locations possibly.