Technical Report on NHS Diagnostic Analysis Using Python

Introduction

Our team was tasked by the NHS to perform an analysis on a provided dataset studying utilisation of NHS services and more specifically we were asked to explain reasons for people missing appointments. As reducing missed appointments would be beneficial financially as well as socially, we keep our focus on replying following questions:

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

Initial Phase of Analysis

We have been provided with dataset comprised with 3 subsets and metadata describing explanation of the studied fields as well as the quality of the data. We advanced our analysis on the premise that data provided was partially clean from duplications, zero values and other initial needed cleaning (which also proved to be the case as confirmed during our analysis). Our initial data sensing basis:

1) **actual_duration.csv** (imported for data wrangling as "`ad`" dataframe). As metadata also confirms, the dataset provides historical data, from **December 2021** until **June 2022**, among other remarks we noted:
   - sub-ICB location level so that we can associate busiest NHS centres with our regional
   - appointments count per day (we can utilize timeseries to check trend)
   - actual duration of the appointment (which helped us recognize trends on busy and non-NHS centres)
   - the subset has no zero values in its rows
   - the column "`appointment_date`" as object which we need to transform to get better plot and understanding for time series analysis
   - Descriptive statistics show the daily mean "`count_of_appointments`" to average around 1,200 for the whole dataset with max prices 10ply higher like 15,400 which is usually found in the busiest NHS locations like London or other big busy city hubs.
2) **appointment_regional.csv** (imported as "ar" dataframe). The range period of this dataframe is different than the actual duration as we start from **January 2020** but ending in **June 2022**. Our remarks include:
   - "`icb_ons_code`" field useful for association with the other datasets on NHS that produce biggest and lowest traffic
   - count of appointments is aggregated per month and also described by categories like appointment attendance status, and others useful fields that help us in following steps draw conclusions or at least recognize trends.

- subset presents no zero values in its rows though data (as also described in our metadata file) has sufficient of the categorical status (mode, status, type) belonging categories like inconsistent mapping.
- View to the monthly aggregation of the " `count_of_appointments`" the descriptive on the datasets show a median monthly of about 1,240 appointments with max values (which are not considered outliers but valid representation of our dataset) which reach 200ply higher figures (ie 211,265)

3) **national_categories.xlsx** (imported as "nc" dataframe). The range period for this dataframe starts in **August 2021** and ends in **June 2022**. Noticeable remarks:
   - "`icb_ons_code`" field also present to proceed to association with busiest or slowest NHS centres
   - Count of appointments offered on daily basis which helps plotting timeseries and also data type formatted correctly as datetime
   - Other categorical fields as areas of service helped identify patterns and trends though quite significant amount of data associated with NA types of category.
   - Descriptive function also shows pattern of max values 10ply bigger (16,590) than the mean daily average of 1, 084.

Analysis and Trends identified

We have identified that the data quality of the provided data set is far from ideal and what would help for a future similar project is to gather historical data for the 3 given subset sharing same period range but also appointments sum on a daily basis. That way would have more leeway in merging dataframes and identifying more patterns on the underlying problem of nonattendance.

Nevertheless, we will summarise our results by answering following questions:

- **What is the number of locations, service settings, context types, national categories, and appointment statuses in the data sets?**

  We have identified **106 different NHS locations** among which we drew the top five with the highest records and the bottom 5. The differences in numbers is astonishing but we comprehend that busiest NHS locations have also the personnel to handle the traffic. Our analysis was focuses on the following categories:
  - 5 service settings
  - 3 context types
  - 18 national categories
  - 3 appointment statuses

- **What is the date range of the provided data sets, and which service settings reported the most appointments for a specific period?**

  Exploring the ad dataframe we identified that **appointments were scheduled from 2021-12-01 00:00:00 until 2022-06-30 00:00:00** but the nc dataframe contains appointments between **2021-08-01 00:00:00 and the maximum is 2022-06-30 00:00:00**

- **What is the number of appointments and records per month?**

  Focusing as a sample on one of the busiest locations (NHS North West London) the service of **'General Practical'** gathers the biggest volume of appointments with a total volume of **4.8 millions of appointments between** the month of **January and June 2022** which averages **about 800,000 appointments per month**. Further investigation shows that overall the busiest month was **November 2021** with a mere total of **30,400,000 appointments**

- **What monthly and seasonal trends are evident, based on the number of appointments for service settings, context types, and national categories?**

  The following diagram (*Figure 1*) shows clearly the difference in volume concentrated in the General Practice compared to the rest of the service settings. All four service settings combined cannot reach the one fifth of the appointments handled by the GP Setting. The line shows a relative increase during autumn months which further declines to stabilize for some months in winter and then a seasonal pattern of ups and downs but keeping stable the average volume above 20,000,000 (figures which as explained before are probably pushed upwards from the big volumes of NHS locations in city hubs)

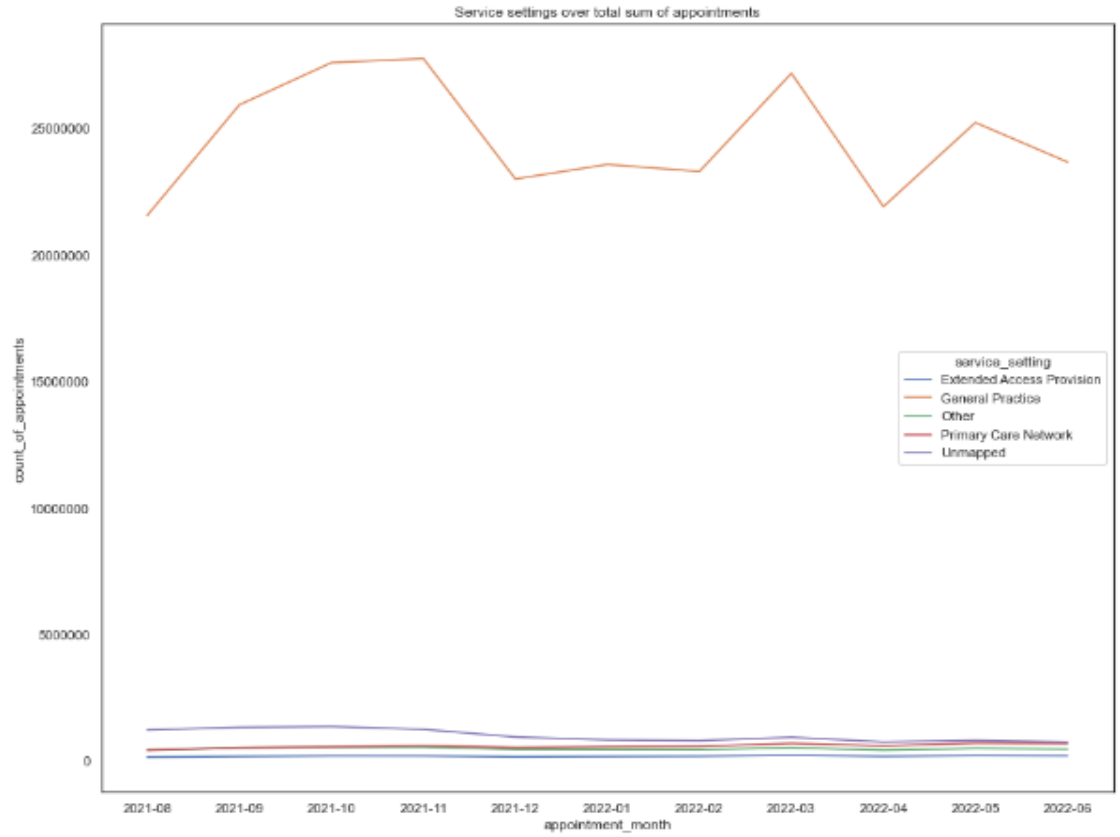


Figure 1 Service Settings total sum of appoinments

We aggregated the appointments per month and determine the sum of the appointments per month over the 3 different context types. The diagram (see

references section) shows same pattern as above and declares a clear winner in the **most occupied context** type as **Care Related Encounters**

Same scenario faced when we plotted the national categories over monthly sum of appointments. Due to the **large number of categories (17)** we have plotted a diagram with the **top 7 national categories** (see References) where again the line shows the same seasonal pattern identified earlier and as category **winner** the **General Consultation Routine**. Though during this representation, the rest of categories show volumes comparable to the top ranking and as well category as the **Planned Clinical Procedure** show a **completely different pattern** (a peak of appointment during October 2010 and then a flat tendance for rest of period)

Last but not least, we sampled certain months to represent the different four seasons and check for related patterns (ie Summer- August, Autumn-October, Winter -January, Spring-April). All diagrams produce same pattern as the one displayed below (*Figure 2*) that clearly shows a peak of appointments on first two days of week which slowly decline during the end of week and starts again on Monday.

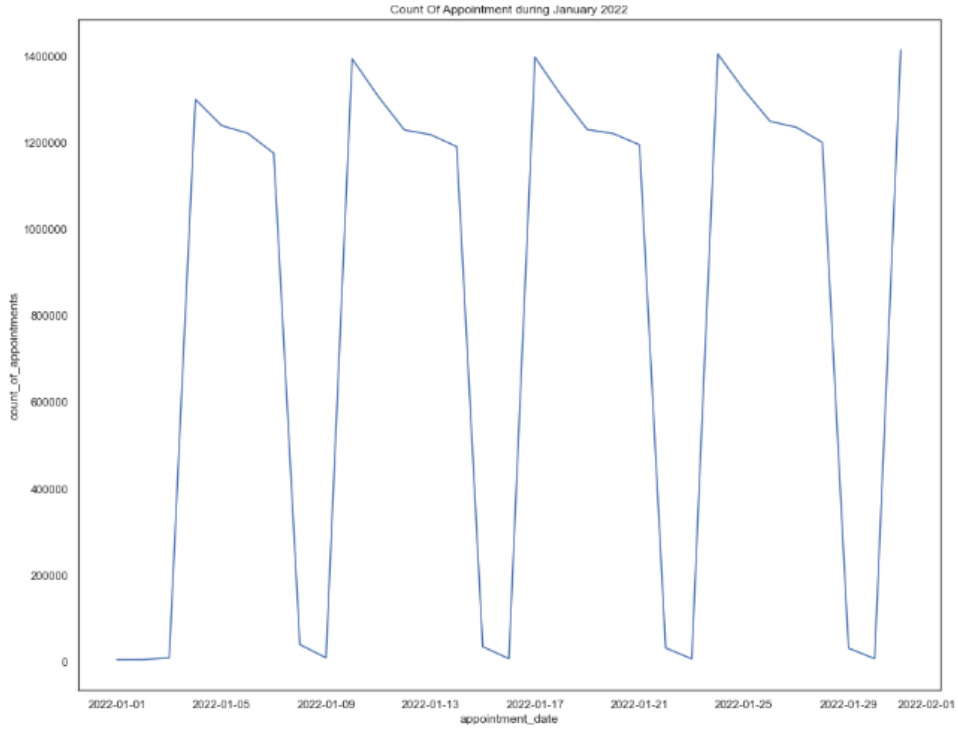


Figure 2 Season Trend during Seasons Analysis

- **What are the top trending hashtags (#) on Twitter related to healthcare in the UK?**

We examined a dataset of tweets with hashtags related to healthcare in the UK. Some very useful conclusion drawn were that :

i) When analysed the retweeted tweets and the favoured tweets one would expect that the figures from those columns would go along. Though what is

also established in the social media universe is that trends don't always show the real image of the case. Something that may be forwarded/shared multiple times doesn't mean it is also favoured as much at the same time. The famous saying :"There is no such thing as bad publicity" unfortunately doesn't have a place in the healthcare industry.

ii) we can see that using the hashtag "healthcare" will guarantee viewing of our tweeter campaign as it is encountered with very big difference from second hashtag in place. In order to make our plot more readable and view that results with count over 10 were not closely related to our study, we also produced a graph basis the top 15 used hashtag from our twitter feeds analysis. (see References)

- **Were there adequate staff and capacity in the networks?**

In order to reply the above question, we concentrated on the `ar` dataframe with date after August 2021 . Displaying results (see *Figure 3*) can clearly identify that max utilization of the services reach 1,013,500 appointments out of a capacify as stated from NHS of 1,200,000. Above comparison leads to the result that capacity and extra staff is not the solution to combat no show attendance in appointments.

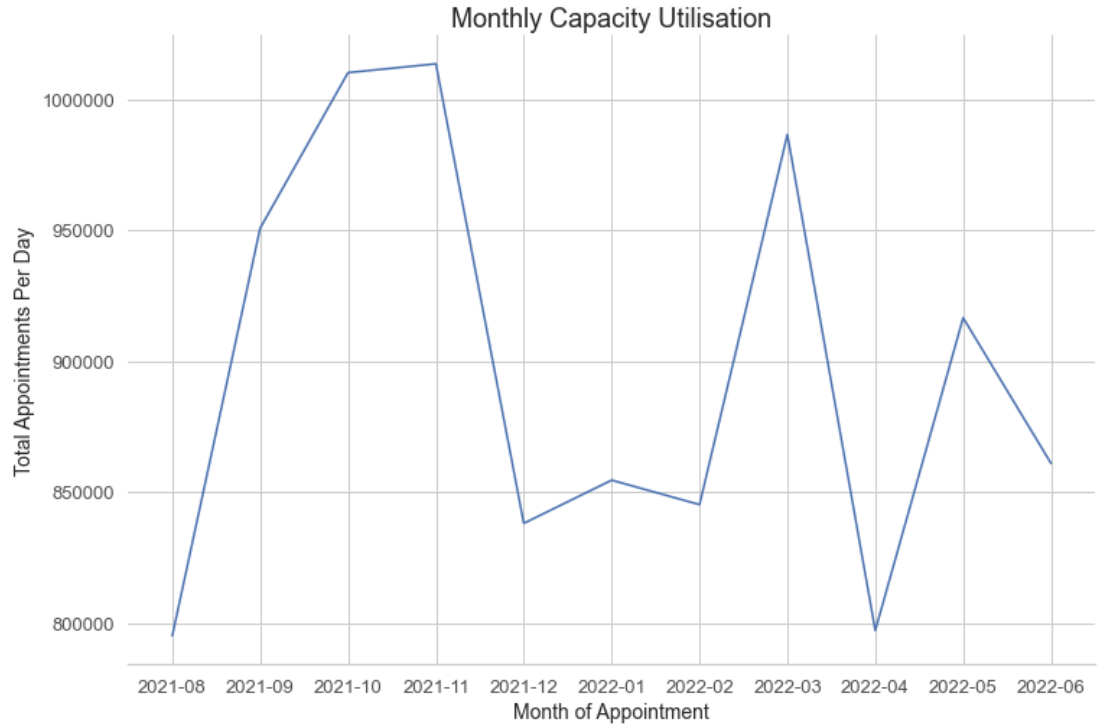


Figure 3 Monthly Capacity Utilisation

- **What type of services are people after?**

We explored the healthcare professional type to identify the sector that NHS should focus its attention on and results of time showed that as expected the General Practitioners assume the majority of the workload (see *Figure 4*)

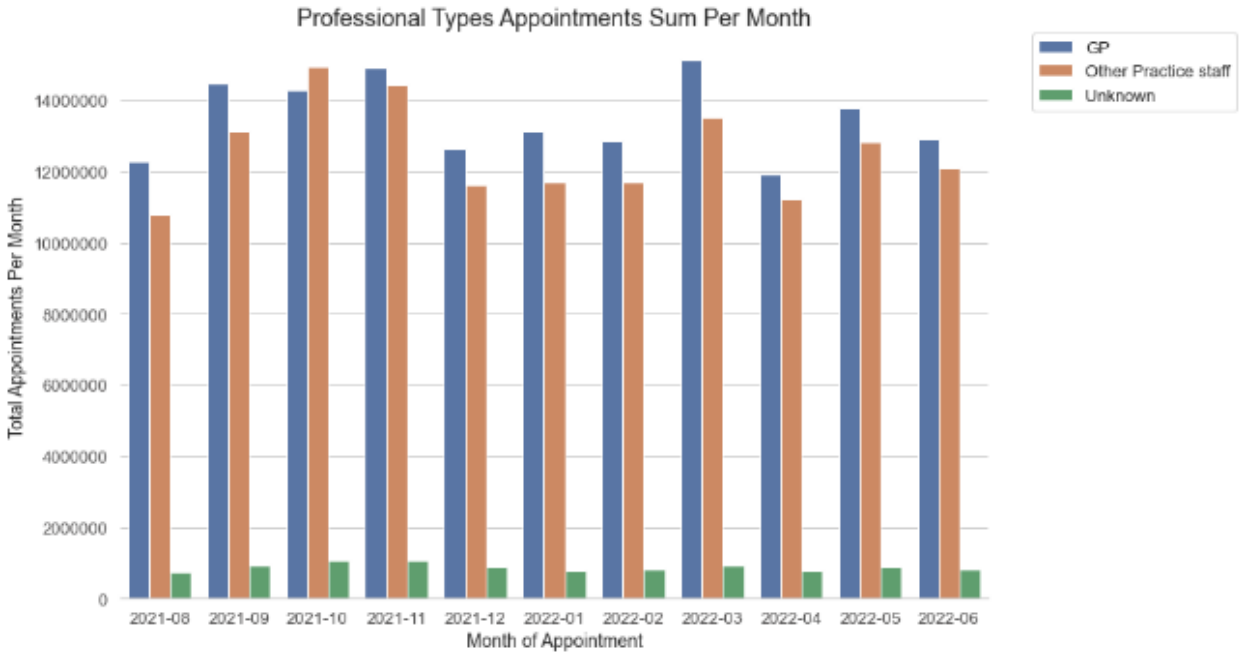


Figure 4 Professional Types Utilisation

One of the flaming questions in the analysis is whether the appointments booked are indeed attended which is the case as show form following diagram (see *Figure 5*)

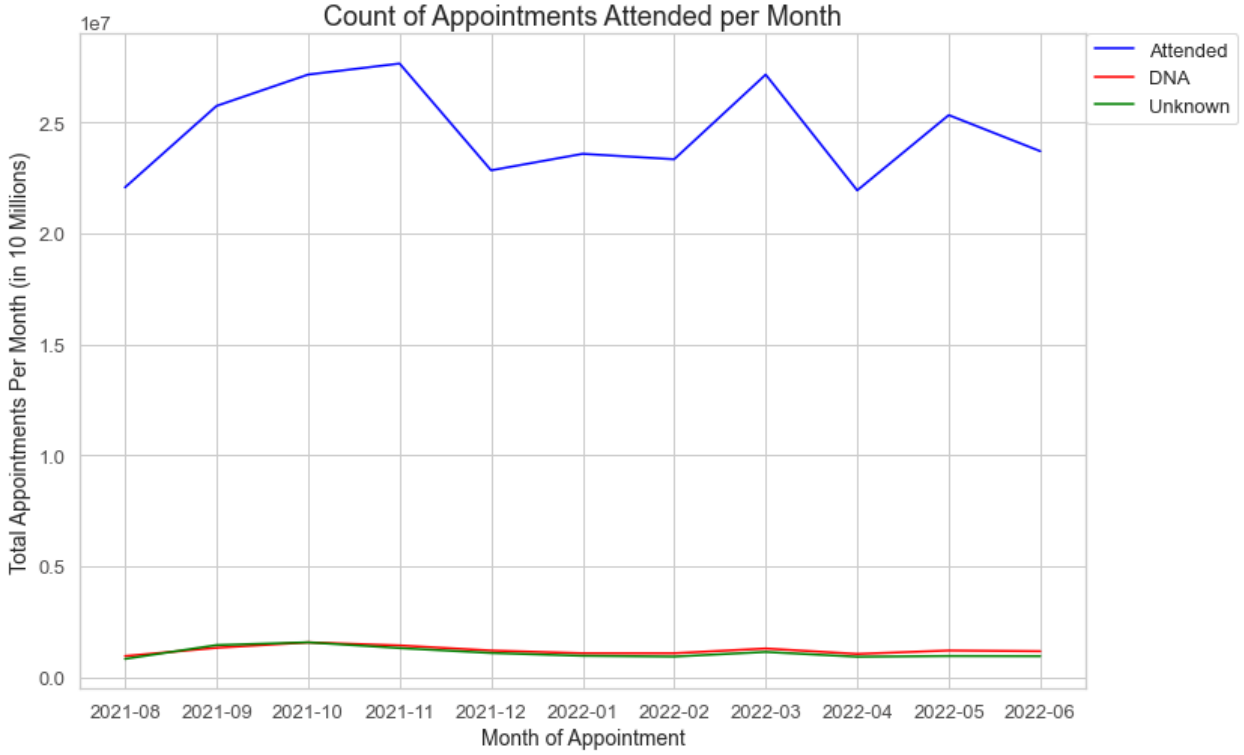


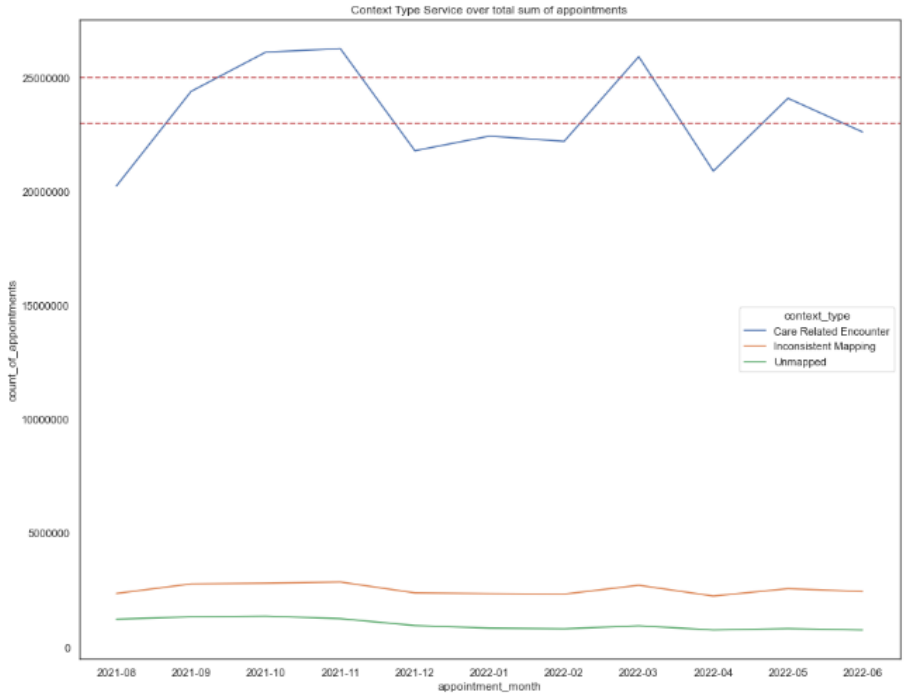
Figure 5 Attendance of Appointments

Conclusion

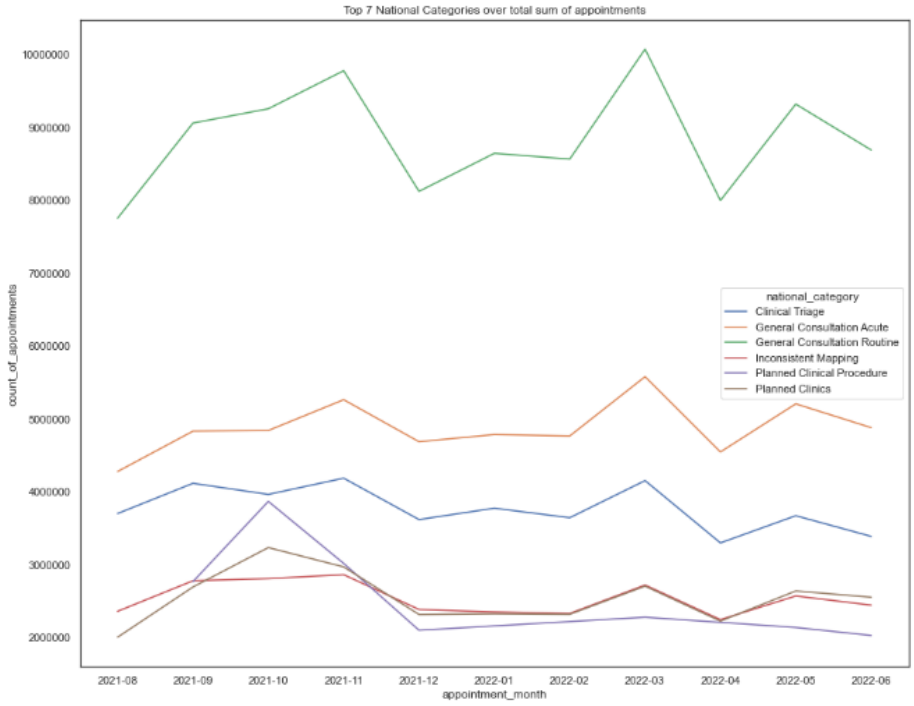
Study of the underlying data gave us the argument to support that the system is working within its designed tolerable capacity which still has a margin of about 200,000 appointments per day (nation wide that is) before reaching its limits. The different data ranges studied from the dataset could not assist us in merging some dataframes under same month, for example or same NHS centre in order to compare all parameters under same range. Though we took a pretty good idea that there is a peak of traffic on the first days of week and also in some cases during certain peak months. This peak is concentrated on one specific type as everybody is after GP services and in the busiest hub cities. We would recommend a re-shuffle of scheduling to assist in spreading better the big volume of appointments and possibly bigger utilisation of the weekend slots that seem empty. The fact also that the majority of the appointments target the GP could may be solved by NHS re-scheduling their appointments forwarding ensuring that other Health Care Practioners working in the NHS could assist in the better handing and scheduling of generated traffic.

References

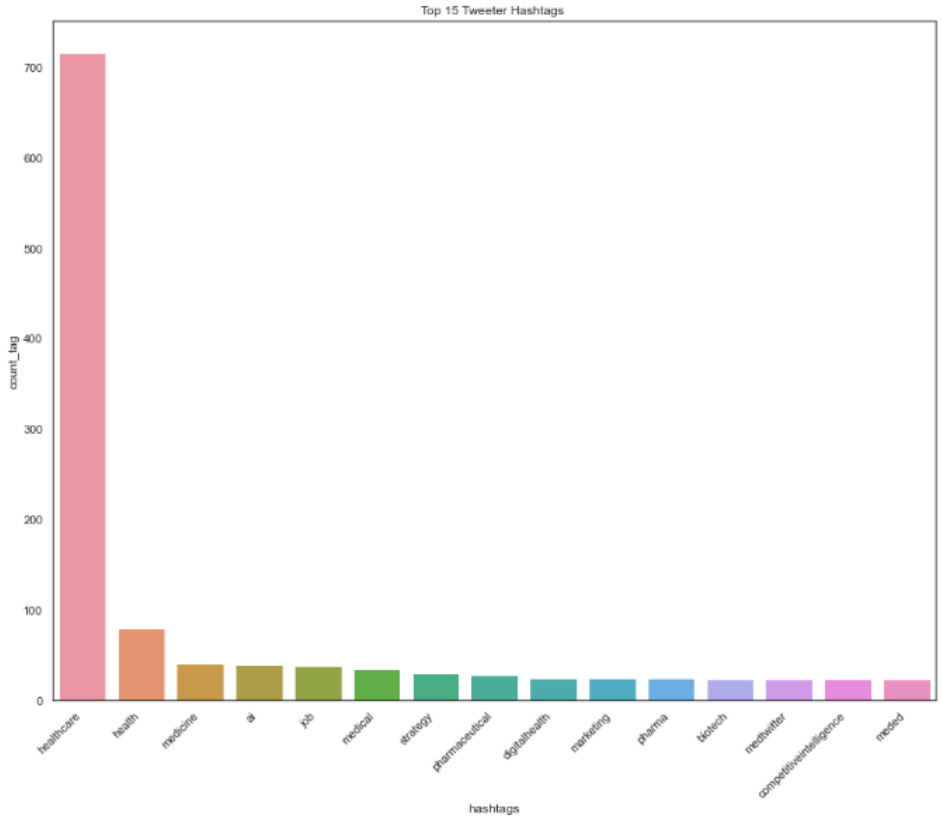
1. Context types diagram



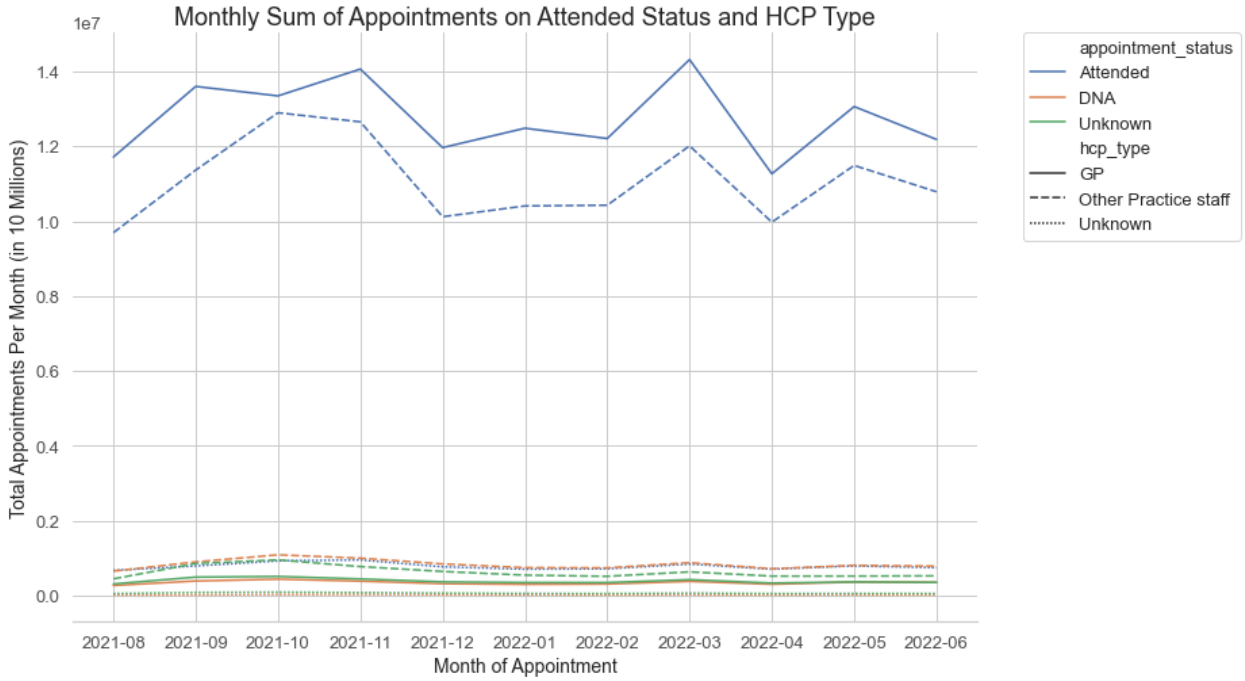
2. National Categories Diagram



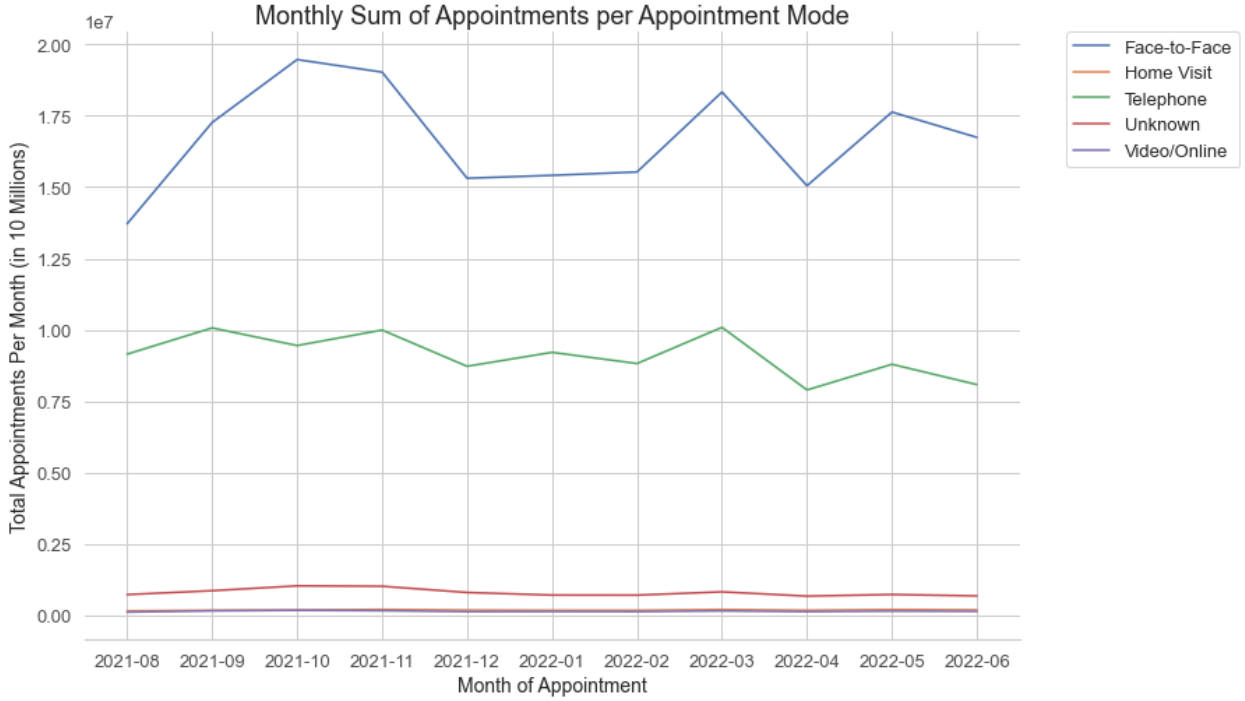
3. Twitter Diagram



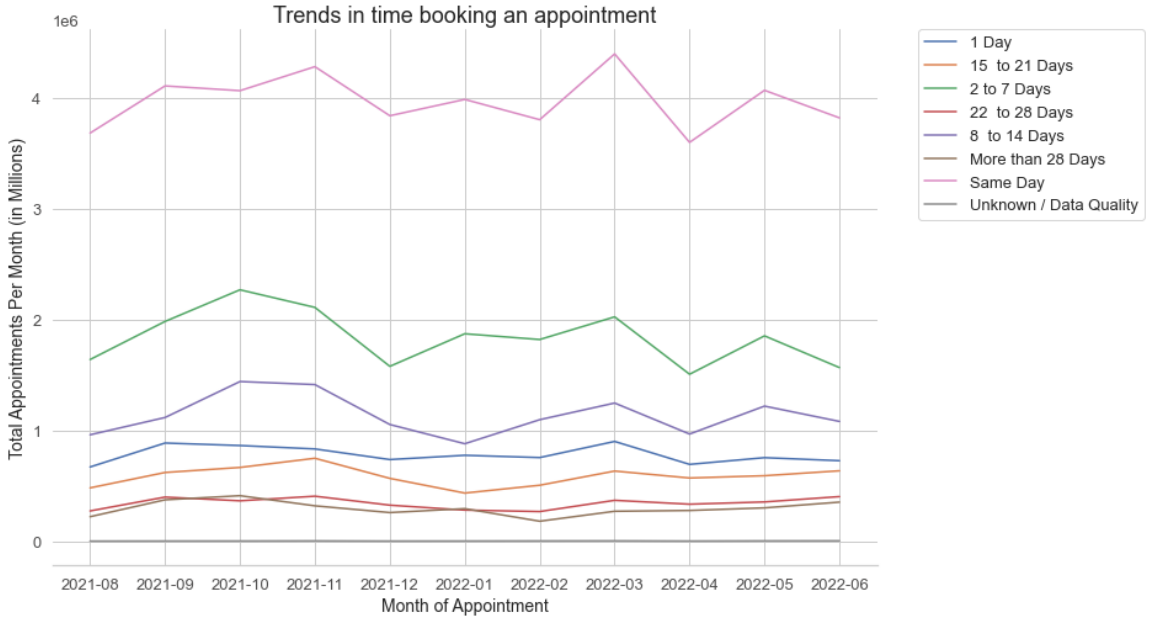
4. Montly sum of Appointments on Attended Status & HCP Type



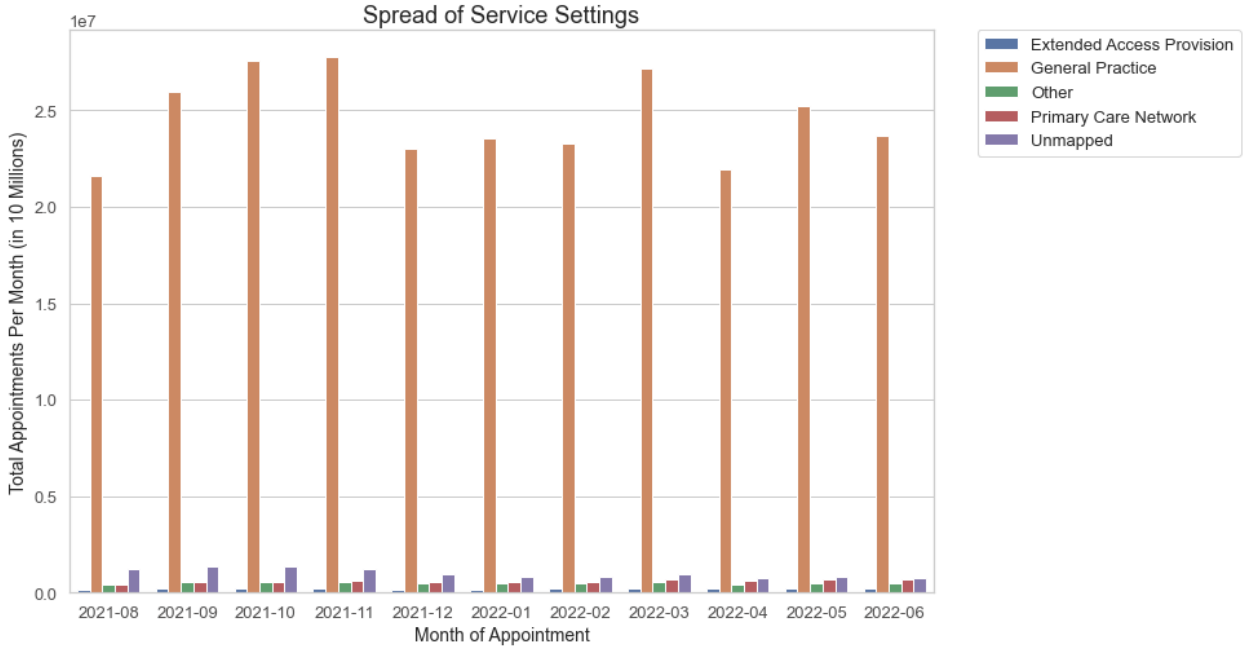
5. Montly Sum of Appointments per Appointment Mode
(**observed same pattern as monthly appointments described in report**)



6. Trends in time booking an appointment



7. Spread of Service Settings



8. Boxplot of attendance (showing skew increasing appointment)