London School of Economics

Data Analytics Accelerator Programme – Summer Cohort CO3

Course:	CO3 Data Analytics Using Python
Assignment:	Diagnostic Analysis Using Python
Prepared by:	Christos Pieris
Date:	02 November 2022 (Approved Extension)

Content of this report

Submitted deliverables in GitHub repository

1. A Jupyter Notebook (IPYNB)

2. A recording (as a .mp4 file)

3. A PDF report indicating the approach, thought process, results, conclusions, and recommendations (pdf)

1 Background and context of the project

1.1 Project Background

National Health System or NHS is UK’s publicly funded healthcare system and is facing several issues now. One of them is the economic and social cost of missed appointments by GP patients. This results in inefficient resource allocation, and it is potentially an area improvement since the reasons for missed appointments are not entirely understood.

1.2 Problem Statement

This project has been contracted to help NHS derive insights from their existing data that will help them understand this issue by answering these two questions:

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

1.3 Scope of investigation covered in this report

While the task can be observed by multiple angles, this investigation considered a range of key items which are summarised below and detailed in the rest of the report.

a. Size, categories, and date ranges of the available data
b. Initial / basic statistical analysis on the prominent categories
c. Monthly and seasonal trends derived from the available data
d. Top trending NHS related tags on Twitter
e. Possible recommendations derived from available data

2 Analytical Approach

2.1 Project development environment:

We are using Python as the language to import, manipulate and interpret the available data. Python has been selected for its powerful libraries yet simplistic almost intuitive nature.

The environment is Junyper Notebook; a web-based interface that allows development on the fly, bypassing the need for heavy duty Integrated Development Environments which are often costly and more complex to maintain.

Our collaboration mechanism for development is Gitlab where updates are passed from Junyper notebooks in Gitlab repository branches which merge to the base code allowing team members to continually work on the project without disrupting the continuous improvement of the base code.

The Gitlab repository can be accessed by following the link below: https://github.com/KrispyPi/LSE_DA_NHS_Analytics

2.1.1 A note on the available project data

1 Excel, 3 .csv files and 1 text file provided by the NHS containing features that are dated on daily and monthly level, characterising the various service settings and events such as whether a patient attended or not.

2.2 Processes taken to explore the data

2.2.1 Importing Libraries:

The Pandas and Numpy libraries were first and foremost brought in to perform the core elements of reading and manipulating our data. The Seaborn library was of vital importance in visualising the results of our operations in an easy and reproducible way.

2.2.2 Importing the Data:

The Junyper Notebook was used to run Python commands to import the files, hence the notebook and files shared the same local directory. Pointing to the files was straightforward by utilising the name of the common directory in the Pandas read_csv (or excel in one case) function which was used to import the files. Furthermore, the result of this operation was a Pandas Dataframe, a 2-D table holding the imported data. The theme of using Pandas Dataframes as the medium to import, read, store, and manipulate data was found to be an effective mechanism for the purposes of this project and it was carried forward in the entirety of the project.

2.2.3 Exploring the Data:

With the data imported into data frames, it was possible to start reading into the structure and content of each data set since it important at this initial stage to get an idea of the size and type of our files.

Quickly built rapport with the data sets, by performing the below basic statistics:

1. Printing the first lines to confirm import worked as expected
2. Viewing the information for each data set which gave us the number of columns, rows, and the count of each type of data.

2.3 Processes taken to analyse the data

Further explored and analysed the data by performing the below steps:

1. Identified that our data cover 106 different NHS locations. This was calculated by asking Python to return the number of unique entries against the Location Column.
2. Out of the 106 Locations with the most records were also called out via the value_counts function and are shown below:

```
--------------------------------------------
The 5 top locations with the most records are:
```

```
--------------------------------------------
NHS Norfolk and Waveney ICB - 26A                   1484
NHS Kent and Medway ICB - 91Q                       1484
NHS North West London ICB - W2U3Z                   1484
NHS Bedfordshire Luton and Milton Keynes ICB - M1J4Y    1484
NHS Greater Manchester ICB - 14L                    1484
Name: sub_icb_location_name, dtype: int64
```

3. The count of for each feature as shown below was called out, BUT it was thought to be more useful to transform the number count into a descending percentage of the total count, so that we understand what counts mean in comparison.

```
Service Setting Count:
General Practice                  44.0%
Primary Care Network              22.5%
Other                             17.0%
Extended Access Provision         13.2%
Unmapped                          3.4%
--------------------------------------------
Context Type Count:
--------------------------------------------
Care Related Encounter            85.7%
Inconsistent Mapping              10.9%
Unmapped                          3.4%
--------------------------------------------
National Categories Count:
--------------------------------------------
Inconsistent Mapping              10.9%
General Consultation Routine      10.9%
General Consultation Acute        10.4%
Planned Clinics                   9.4%
Clinical Triage                   9.1%
Planned Clinical Procedure        7.3%
Structured Medication Review      5.4%
Service provided by organisation,
external to the practice          5.3%
Home Visit                        5.1%
Unplanned Clinical Activity
Patient contact during,           3.5%
Unmapped                          3.4%
Care Home Visit                   3.3%
Social Prescribing Service        3.2%
Care Home Needs Assessment &
Personalised Care &
Support Planning                  2.9%
Non-contractual chargeable work   2.6%
Walk-in                           1.7%
Group Consultation,
and Group Education               0.7%
```

4. The dates between appointments were scheduled were called out to see the range of dates under investigation. Firstly, the type of dates was checked to be ‘object’ which was changed into datetime so that normal date-type operations can be performed on these columns.

The min and max dates were then called out resulting in our available date ranges indicated using appropriate docstrings.

```
The minimum date in the ad DataFrame is: 2021-12-01
The maximum date in the ad DataFrame is: 2022-06-30


The minimum date in the nc DataFrame is: 2021-08-01
The maximum date in the nc DataFrame is: 2022-06-30
```

5. The most popular service setting for NHS North West London from 1 January to 1 June 2022 was found by querying the National Categories and filtering the dates between the desired range. Looping through the filtered set each service setting count was shown, resulting in the General Practice being the most popular.

```
The number of records for Unmapped is 829415
The number of records for Other is 313682
The number of records for General Practice is 9473675
The number of records for Extended Access Provision is 204363
The number of records for Primary Care Network is 218658
```

6. The month with the highest number of appointments was found to be November 2021.

7. Finally, the total number of records per month as calculated by grouping the data frame against months and aggregating using the sum of counts found in each month.

3 Visualisation and insights

3.1 Overview

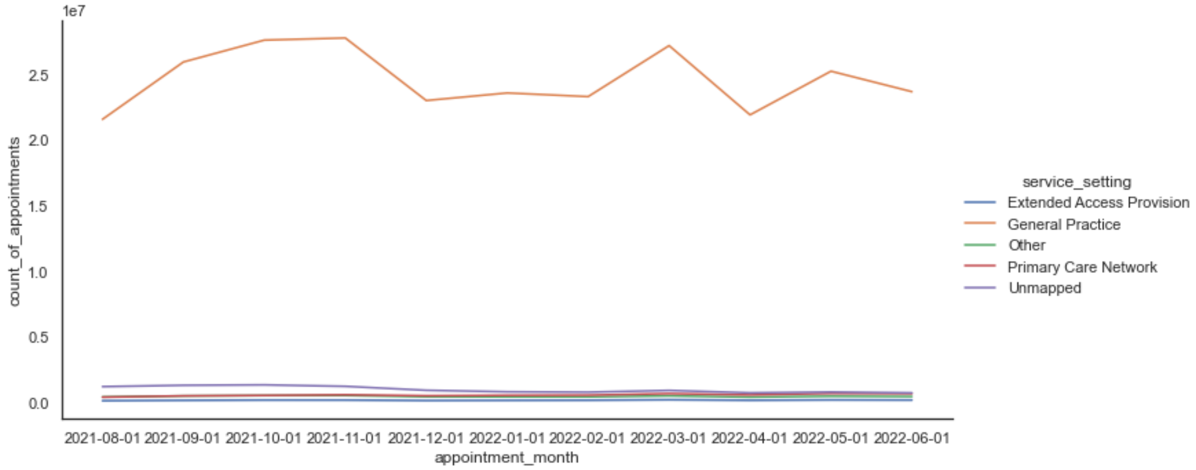
3.1.1 Seaborn library was used and originally imported into the working environment. The chosen library is built on top of the extensive matplotlib library and allows parsing data in Pandas based modules. This was a very meaningful step to easily produce consistent graphical representations of our data. Finally, the figure size, background colour and palette were kept consistent throughout for consistency.

3.1.2 By analysing the data, we identified initial trends, and it was evident that visual representations would help surface them. While the tables that resulted from the analysis provided the facts, it is the interpretation of different facts when put together that this stage achieved.

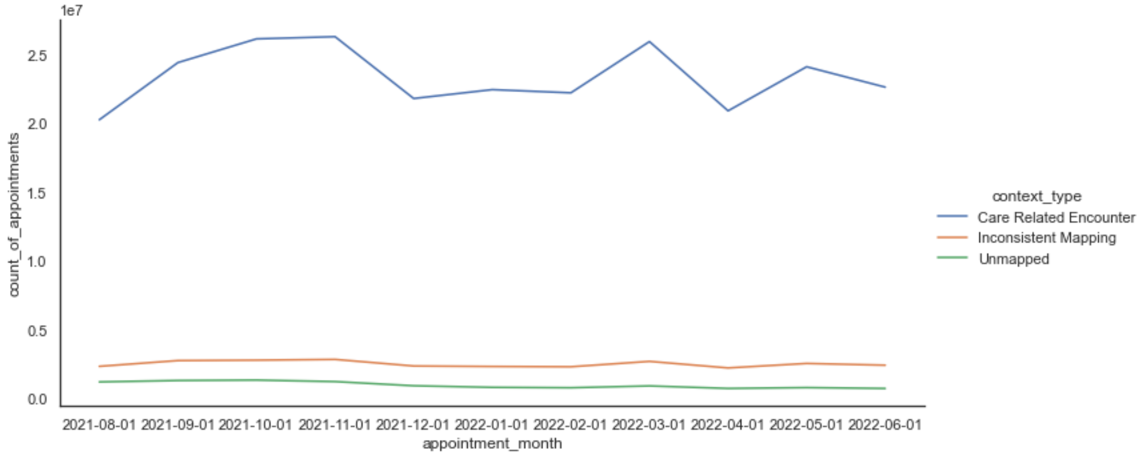
3.1.3 The project utilised mostly Lineplots for continuous data such as the change in service type counts against time. Barplots were also utilised for categorical data such as the count of unique Tweeter hashtags.

3.2 Visualising monthly trends in the data

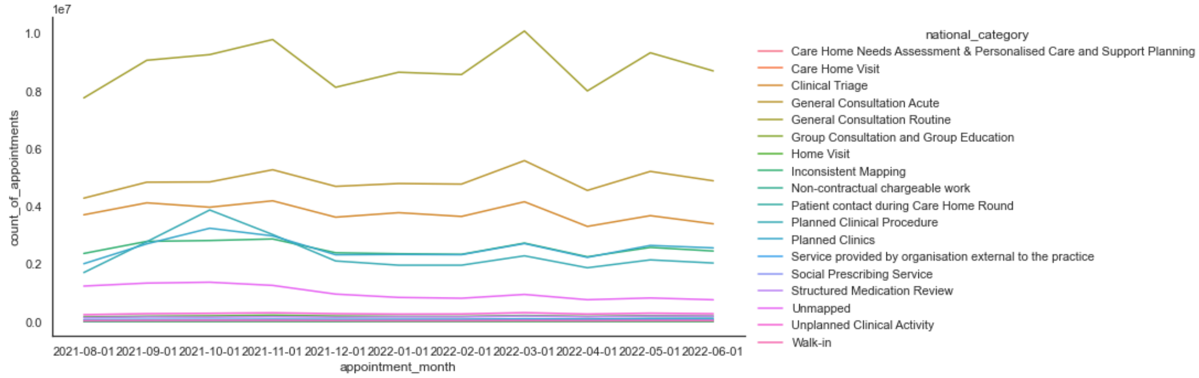
3.2.1 The number of appointments per month for each type of Service Setting



3.2.2 The number of appointments per month for each type of Context

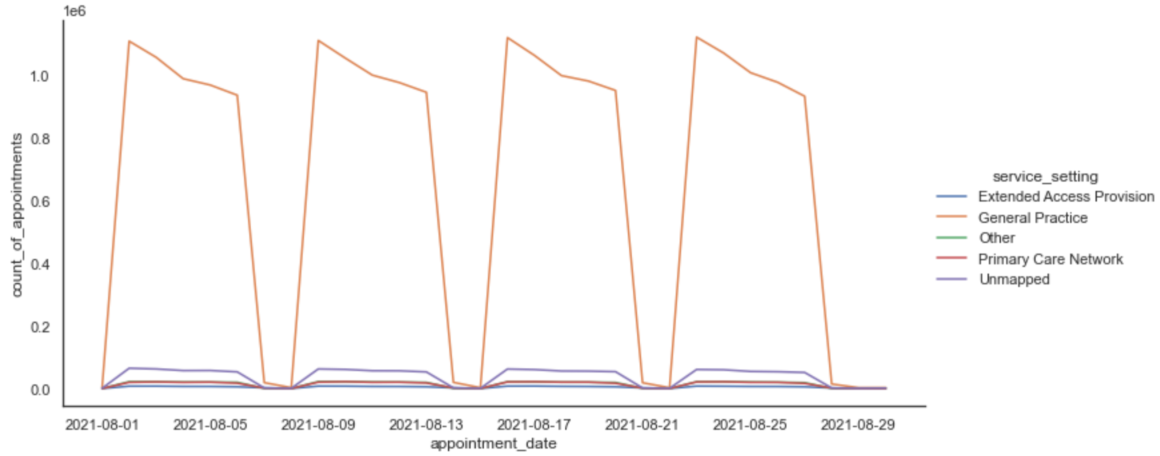


3.2.3 The number of appointments per month for each type of Context

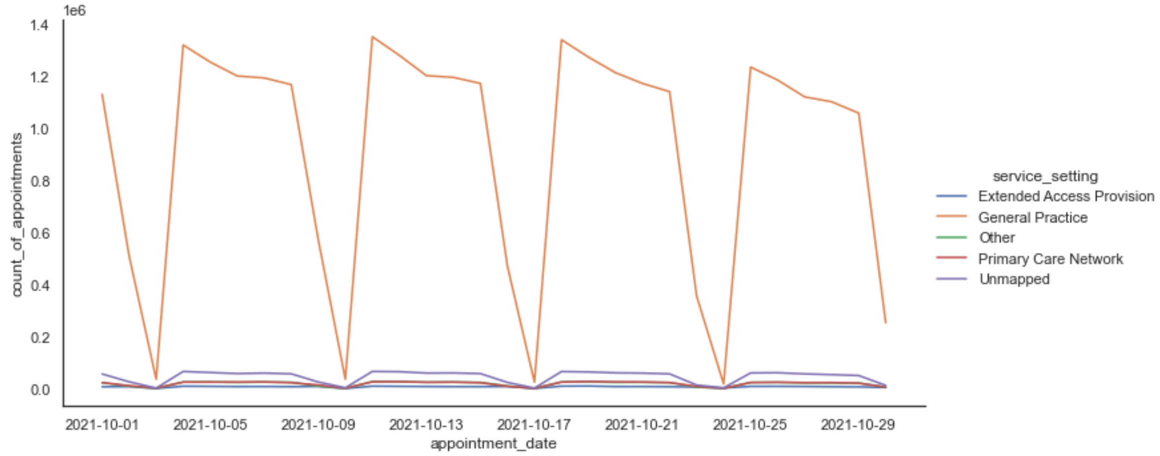


3.3 Visualising seasonal trends for Service Settings in the data

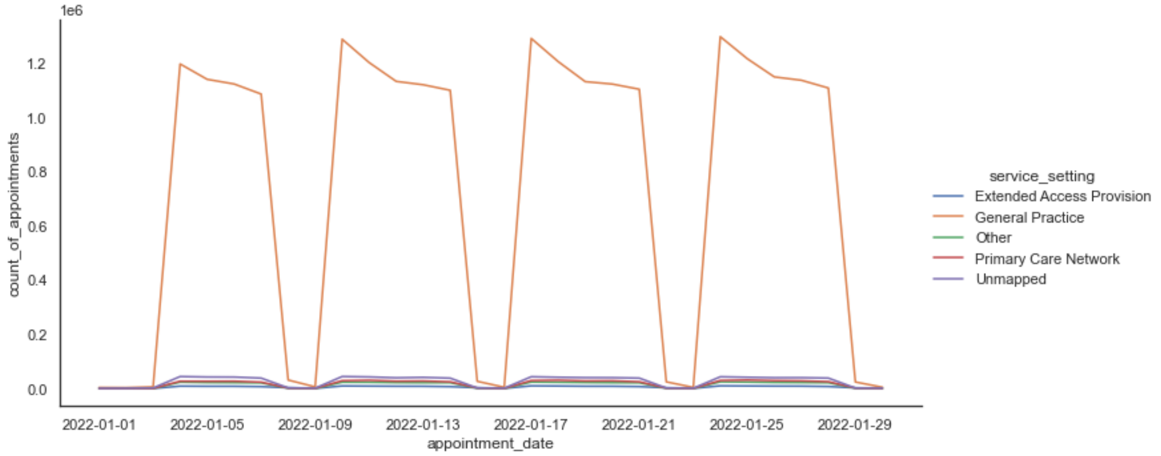
3.3.1 Summer, August 2021



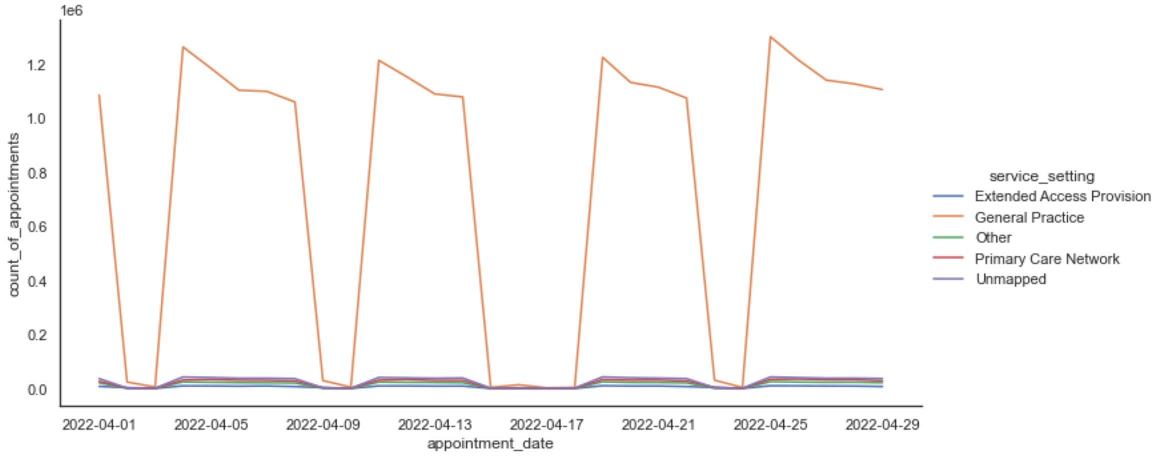
3.3.2 Autumn, October 2021



3.3.3 Winter, January 2021



3.3.4 Spring, April 2021



4 Analysing the data for patterns

4.1 Gauging people's views of NHS on Tweeter

The tweeter data set was introduced to enrich the analysis. Initially the set was imported into Pandas data frame to be viewed, sized and comprehended.

4.1.1 Out of many tweets, few are those which resonate with the user base

By plotting the value counts for the tweets that have been re-broadcasted by users, it was evident that most tweets were in fact left alone once tweeted. That ties in well with the expected notion that many words, phrases or ideas will find their way on the platform, but few will stick out and resonate with the user base.

4.1.2 Choosing to be open to all views

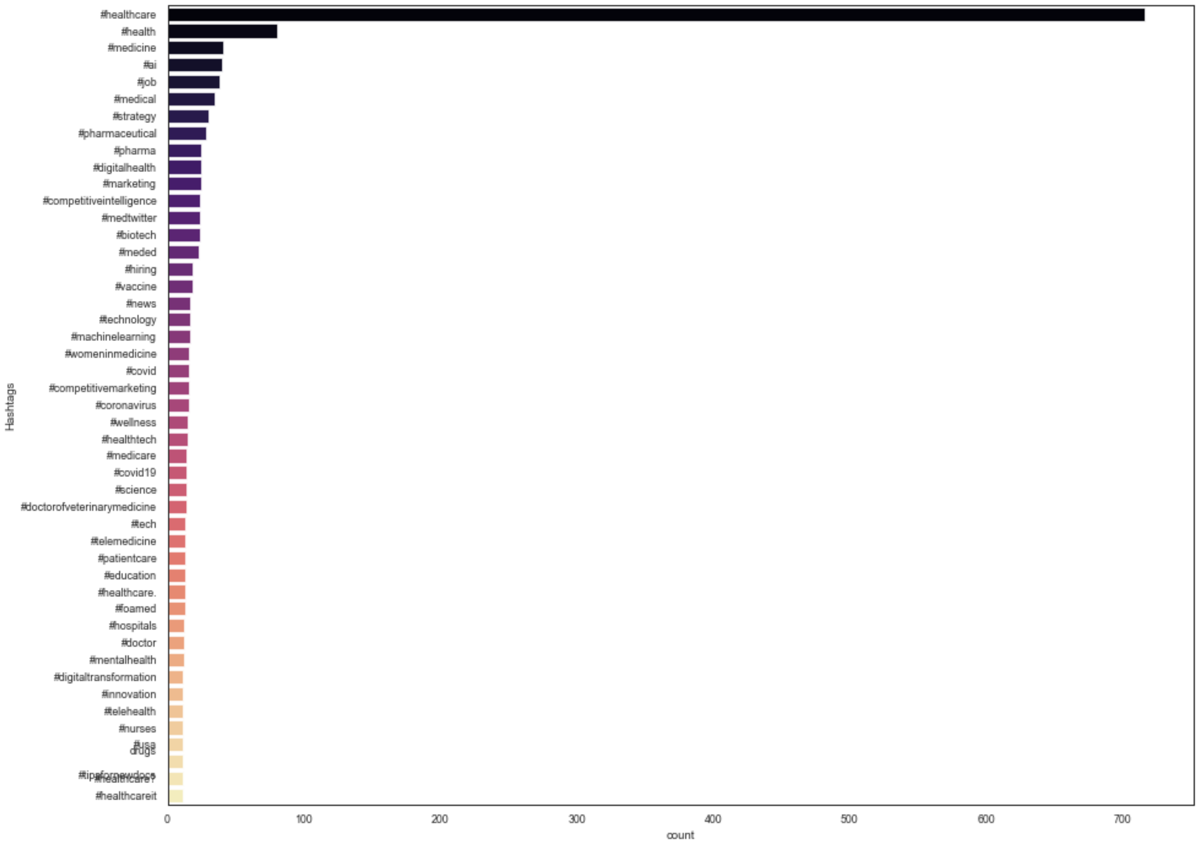
The project decided to continue working on the full dataset to provide the breadth of views expressed on the platform when it comes to NHS.

4.1.3 Extracting meaning out of chaos

The amount of text while viewing the tweets in tabular format was initially daunting. Hence the set was deconstructed by extracting text only and then filtering through each row to extract the hashtags (#) against each tweet. The filtering was performed via a loop that updated a list. The list was then reconstructed as a Pandas data frame, holding each tag and its count. An essential step to be able to easily visualise using Seaborn.

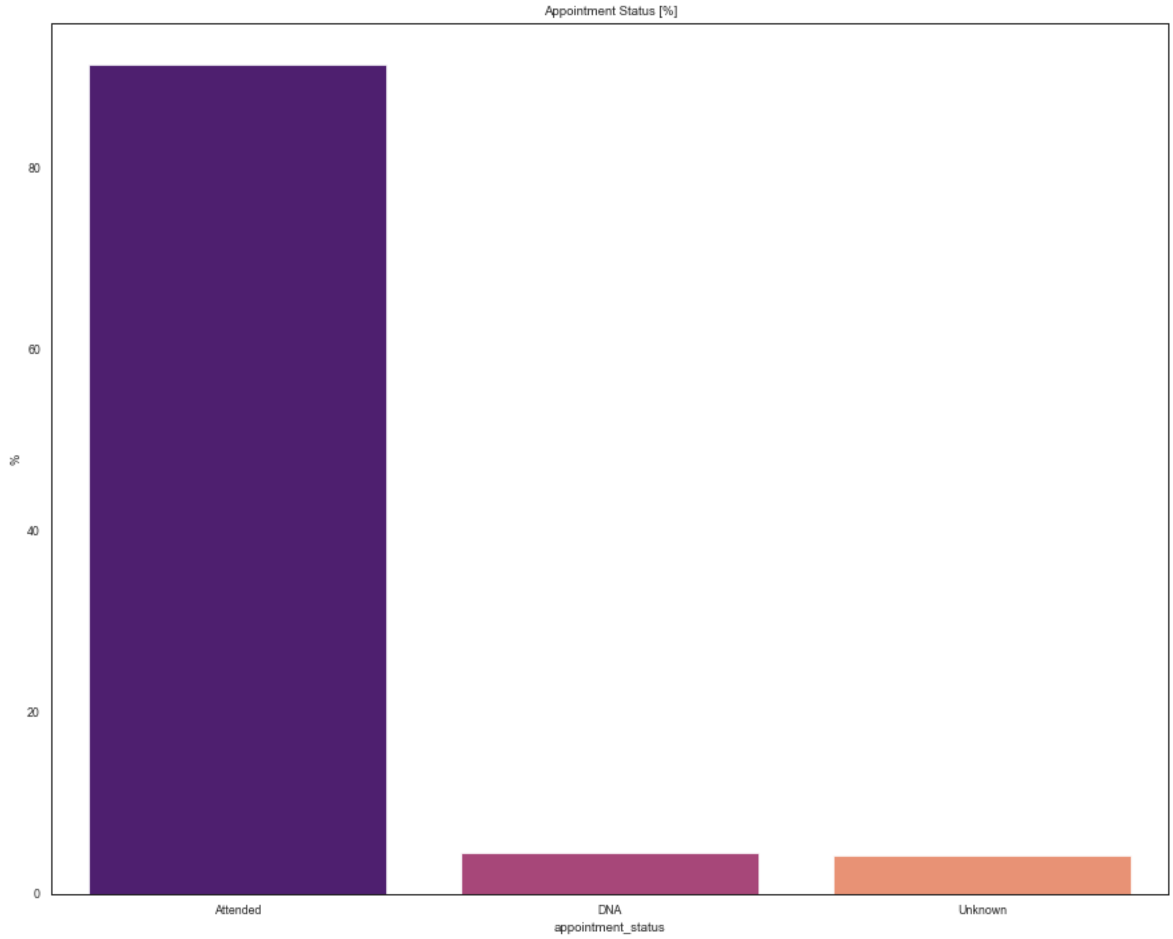
4.1.4 The most used hashtag in tweets referring to NHS is ‘healthcare’, by far

All the hashtags that were found at least 10 times in the set were plotted. A much more meaningful and powerful way to quickly see that the most frequently used tag in NHS related tweets is ‘healthcare’ by far. Followed by ‘health’ and ‘medicine’ unsurprisingly, and then also ‘ai’ and ‘job’ which is surprising.

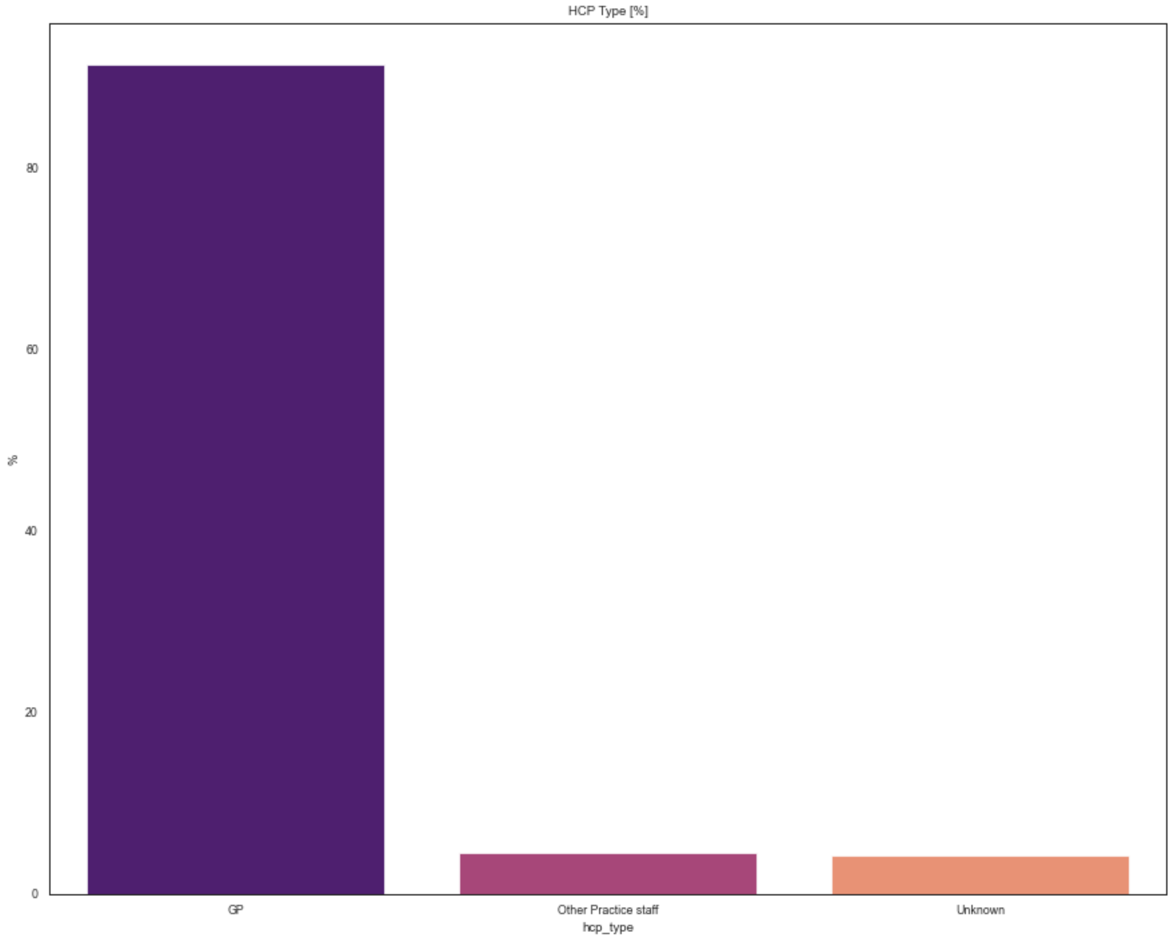


4.2 Gauging the capacity of NHS staff levels

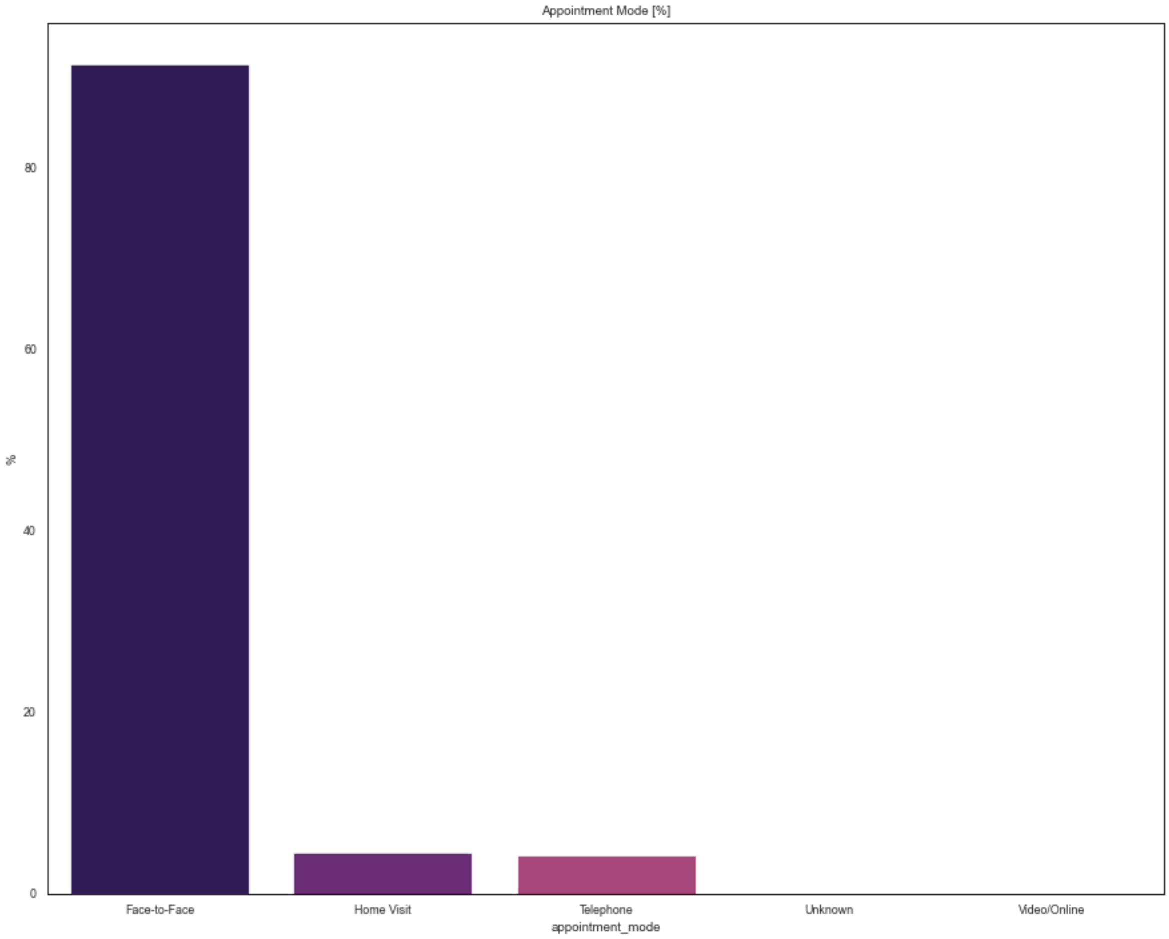
4.2.1 Zooming into the appointment status by visualising the percentage of each category. Most of the appointments in the date range selected attended were attended.



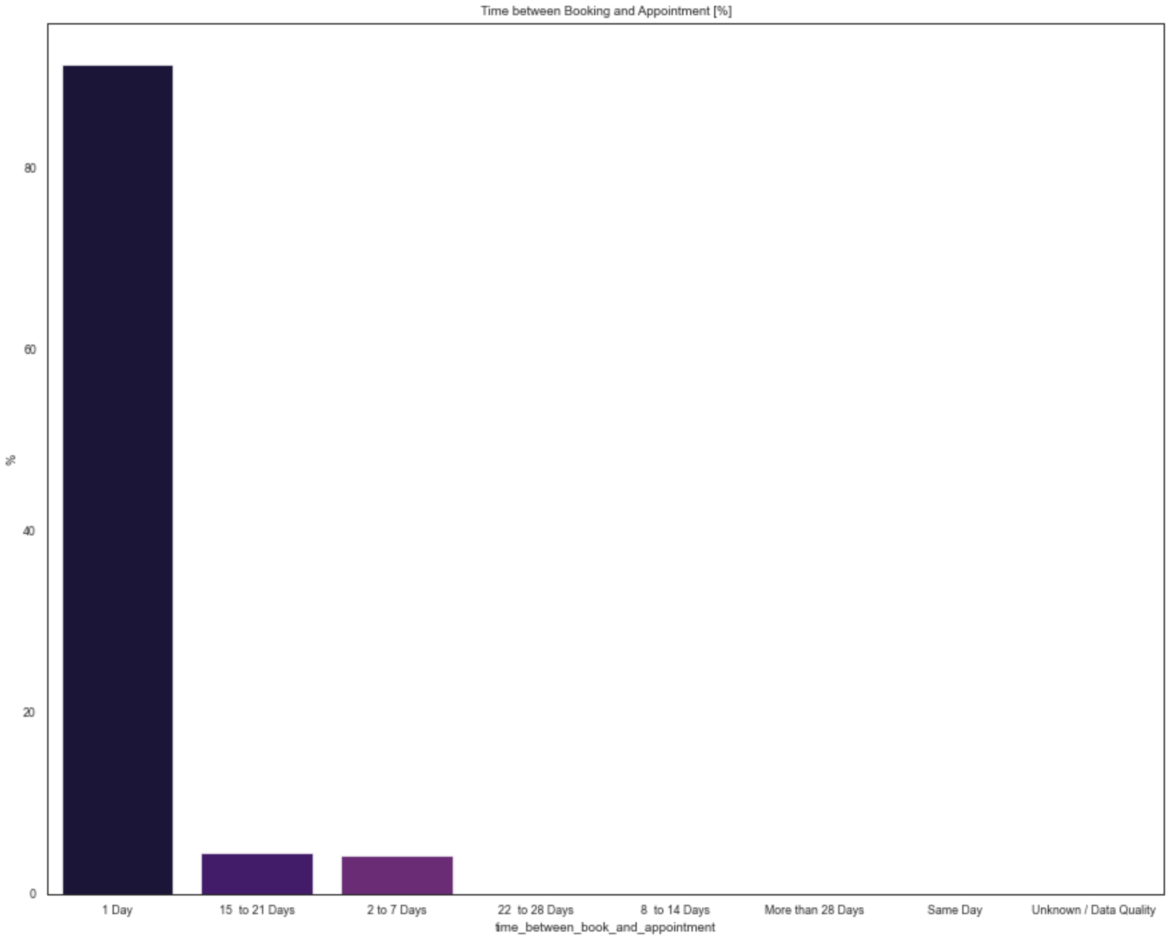
4.2.2 Most of the appointments were met by GPs.



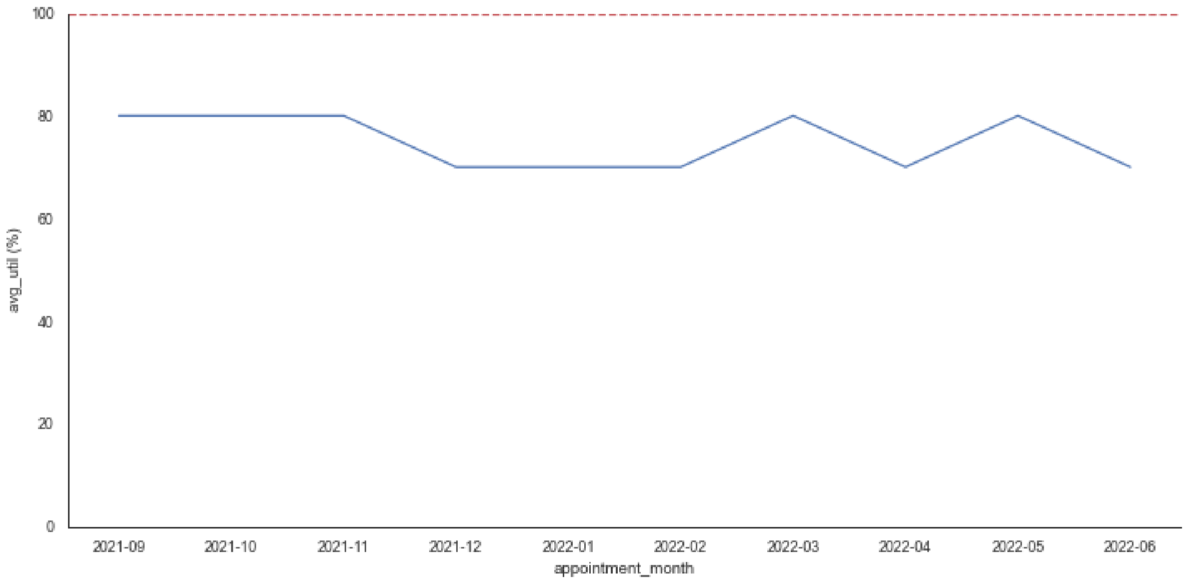
4.2.3 The way that most appointments were performed were face to face, followed



4.2.4 Most appointments were booked a day in advance.

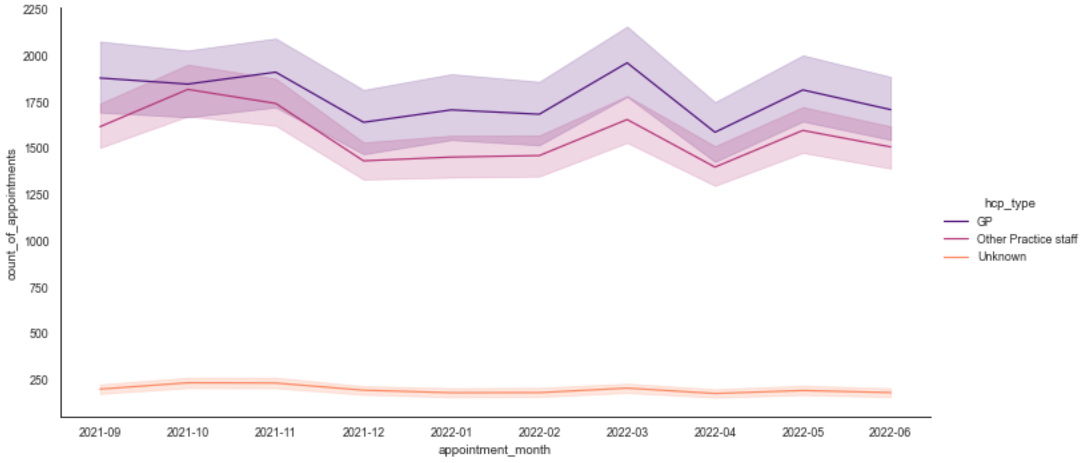


4.2.5 Average utilisation of the NHS service was calculated for each month to range between 70% and 80% with overall average being 75%. This informed the project against the main question posed by NHS, that across the range of dates there was sufficient capacity to meet demand for booked appointments.

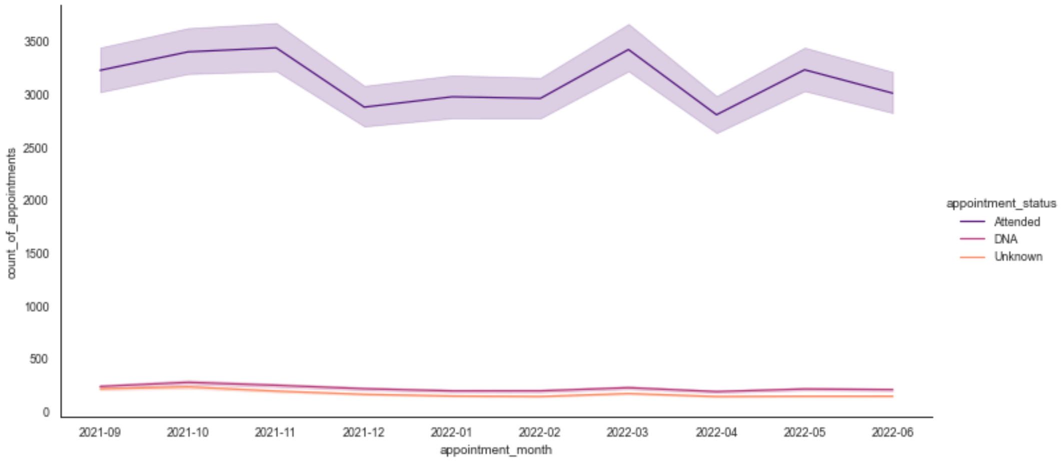


4.3 Zooming into some of the key features

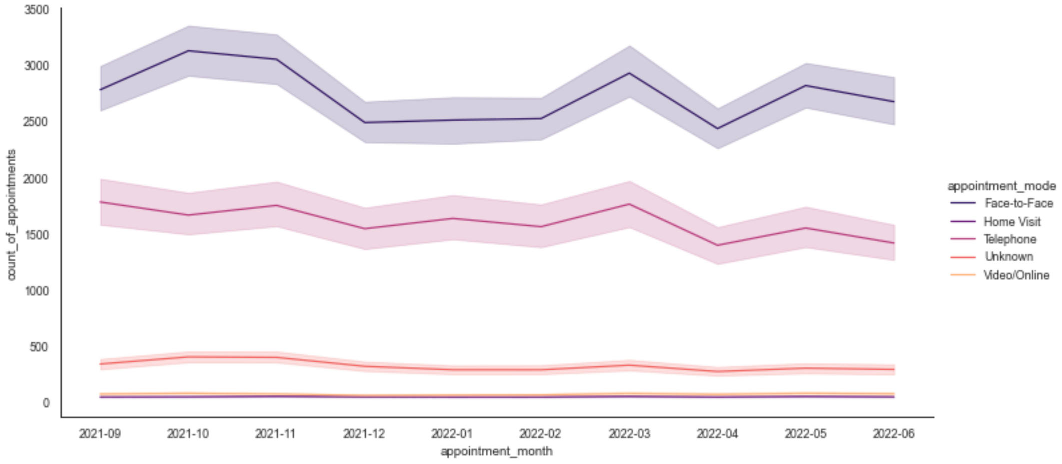
4.3.1 Plotting for healthcare professional types against time



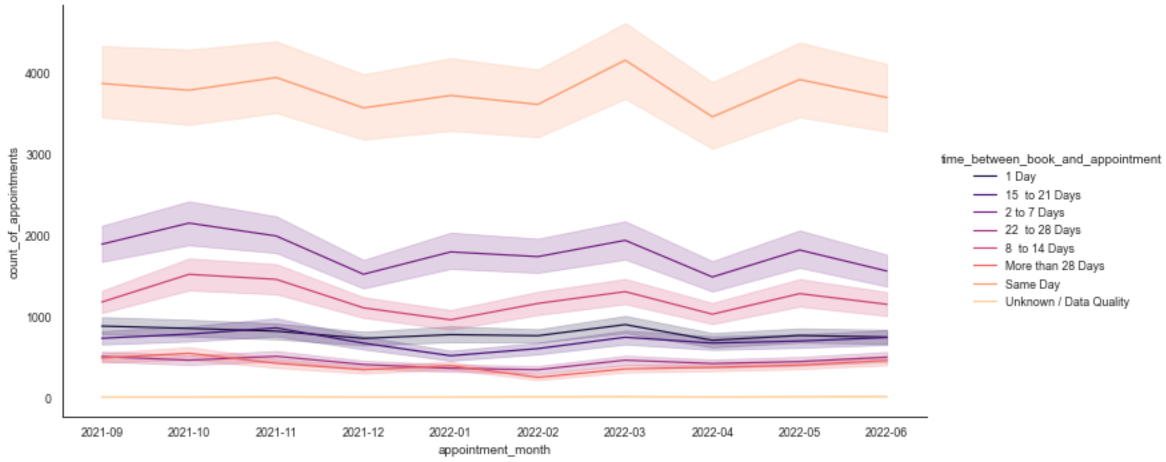
4.3.2 Considering the unattended appointments



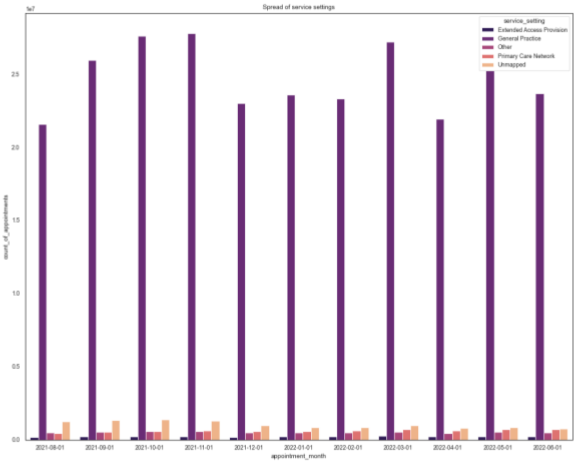
4.3.3 Correlation between appointment type and busiest months

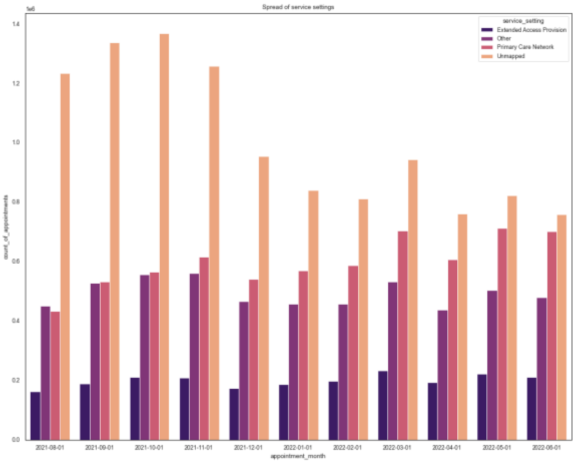


4.3.4 Trends in time between booking and attending an appointment



4.3.5 Zooming into the spread of service settings it was shown that the major service setting was General Practice, while plotting without the dominant category revealed more information on the rest of the data showing that the second largest service setting set was Unmapped.





5 Summary

5.1 Key findings and recommendations

The NHS networks were well resourced to meet demand for the given period, with utilisation averaging 75%. Some of the focus areas should be to mitigate risks where bottlenecks can occur and to refine the current reporting and data management to benefit from data driven operations, especially during the busiest months of October, November.

5.1.1 Analysing historical data covering a period where the UK was affected by the pandemic, the project showed that only a fraction of the appointments booked were not attended (<10%). It is reasonable to predict that this number will only get smaller in the time after the investigated time window, considering the social norm will be restored post-pandemic.

5.1.2 The project identified that the biggest set of appointments are served by GP. Considering this is an area of potential congestion and bottle-neck formation, it is part of this project's recommendation that the NHS investigates for alternative ways to deliver appointments were possible. Investing in upskilling other practice staff and / or improving triage stages could be considered to mitigate this risk.

5.1.3 Face-to-face meetings were shown to be the most popular. The project suggests that NHS invests in other means to deliver consultation and at the same in educating people to take up alternative routes such as tele-health appointments. Resource management will benefit from the agility that virtual settings can offer.

5.1.4 Most appointments were made a day before they were supposed to take place. While this will be a hard requirement for certain health conditions, the project suggests that the NHS promotes longer periods to allow for efficient resource allocation to settle in a more optimal way, given that with more time both the demand and supply will be able to make arrangements and indeed be successful at the point of contact.

5.1.5 The analysis showed that the NHS resource networks were well prepared for the given period. Utilisation never came close to exceeding capacity, indicating that the NHS should focus in optimising the existing resource allocation as well as reducing the missed appointments rather than injecting head count.

5.1.6 Investing in robust reporting and data management is also an area that the NHS can benefit from improving. This was identified by revealing trends in the service setting spread, indicating that the unmapped appointments were the second largest service setting, especially during the busiest months. This means that when busy the NHS reporting suffers, when this is when data driven management will help the most in resource allocation and operational management.

5.1.7 Tweeter is an established form of expression where people seem to refer to NHS in expected ways such as when talking about healthcare, medicine, and health related issues. Analysing the data showed that people include the NHS in other discussions flagged in technology, artificial intelligence, and job-related discussions. This is an area worth taken in by NHS as an input into how the organisation is currently being viewed. This can inform their go – to – market strategies for attracting talent, transforming their operations, and defining as well as implementing their vision.

5.2 Future work

The NHS is invited to extend this project to allow further investigation to take place beyond the pandemic impacted period which is expected to result in more future proof recommendations. Finally, extending the project to include more data sets covering areas in other parts of the network is also expected to result in insights that will better describe the entirety of the NHS network and capability.